TAGGING AND CHUNKING OF CORPORA IN INDIAN LANGUAGES

Rajendran Sankaravelayuthan
Amrita University, Coimbatore
rajushush@gmail,com

1. Introduction

Annotation of corpora is very significant for any kind of NLP oriented researches and applications. Natural language processing, speech recognition and other related areas require annotated corpora to serve as an important tool for investigators. Constructing statistical models for automatic processing of natural languages annotated corpora offers a basic building block. Creation of annotated corpora is the first step towards natural language processing. Annotated corpora proved to be very useful for language processing. So, annotated corpora are created for languages across the world. Unfortunately not much work has been carried out for creation of annotated corpora Indian languages. Unavailability of annotated corpora, large enough to experiment statistical algorithms is the main bottleneck for the computational processing of Indian languages.

Annotation of corpora is done at different levels of language analysis such as part of speech, phrase/clause level, dependency level, etc. The basic step towards building an annotated corpus is Part of speech tagging (POS) and the next level of tagging is Chunking.

The present draft is a compilation of information I have gathered while I was working for Indian Language to Indian Language Machine Translation. The material was lying in my laptop sine long. I thought of making to see the light.

2. POS Tagging

Parts-of-Speech (POS) tagging – the task of assigning appropriate POS tag to every word in a given text – is an important process used as a building block for several Natural Language Processing (NLP) tasks. A POS tagset usually defines the list of morphosyntactic categories that

are applicable at the word-level to a specific language, though in some cases they may also include pure morphological or syntactic information. The early efforts in POS tagset design in 1970s, that resulted in tagsets such as UPenn, Brown and C5, mainly focused on tagsets for English which were mostly simple lists of tags corresponding to the morphosyntactic features, and varied greatly in terms of their granularity (Hardie, 2004).

CLAWS2 tagset (Sartoni, 1987) marked an important change in the structure of POS tagsets from a flat structure with unitary tags to a hierarchical structure that allowed for decomposable tags. This enabled the provision of distinct encoding for all word classes with distinct grammatical behaviour and a systematic approach in building the final tag from its constituent. The publication of EAGLES recommendations for morphosyntactic annotation of corpora (Leech and Wilson, 1996) was an earliest attempt to develop a common tagset guideline for several European languages. The objective of EAGLES1 was to standardise the tagsets used in different projects and/or different languages to achieve *cross-linguistic compatibility*, *reusability* and *interchangeability*. According to Leech and Wilson (1999), at the cross-linguistic level annotations used for one language should as far as possible be "compatible" with annotations used for another. This means that any common descriptive categories between different languages should be encoded with the same string so as to be recoverable from the annotations applied to texts in different languages.

2.1. EAGLES Guide Lines for Morphosyntactic Annotation

The EAGLES guidelines for morphosyntactic annotation are structured on three levels:

1) What is considered obligatory;
2) What is recommended;
3) Optional extensions for properties that are language-specific or marginal to morphosyntactic annotation.

At each level, tags are defined as morphosyntactic attribute-value pairs (e.g. Gender is an attribute that can have values Masculine, Feminine Or Neuter in the EAGLES recommendations). These attribute-value pairs may be structured as a hierarchy but need not be.

2.2.. Indic Language Tagsets – Current Scenario

There are several tagsets used in Indian languages, which have been developed by the research community. While some of these tagsets are sufficient for basic POS tagging, they do not satisfy the requirements of some tools developed elsewhere, which demand finer distinctions and richer corpus annotation.

IIIT-Hyderabad is leading an effort, jointly with several other research institutions to standardize the tagset for Indian languages. While, the effort is still ongoing, the research community has already released the first version of the comprehensive tagset, which has about 27 tags. It should be noted that this is a common tagset, for all Indian languages.

CIIL-Mysore has developed a Hindi tagset with about 31 tags. While this has several tags in common with the abovementioned tagset, it also has a few additional tags to achieve a higher level of granularity.

AU-KBC Research Centre, Chennai has developed a tagset for Tamil, after extensive consultations with several linguists. This has about 64 tags, and achieves a high degree of finer distinctions of the language.

For ILCI project we intends to consider these tagsets as a basis to design our own tagsets for Indic languages.

2.3. Tagset Design – Principles

While, a common tagset for all Indian languages would be better from the point of view of baseline, creation of annotated data etc., the syntactic differences between the languages should also be taken into consideration during the tagset design.

A parallel can be drawn with the European languages, which while sharing a large number of morpho-syntactic features have different tagsets. Monachini and Calzolari (94) suggest that, it would be impossible to have identical tagsets across languages and advocate the need for preserving language specific peculiarities.

We can start with two different tagsets for Indic languages- one for the Indo-Aryan family (most North Indian languages) and another for Dravidian family (most South Indian languages). This is based on the fact that the two language families are significantly different in their morpho-syntactic behaviour. However, we must be open to suggestions on any special tag(s) that will be required to capture specific features of individual languages as extensions.

Another important question in tagset design is the type of tagset- a *simple list* or *hierarchical*. While the simple tagset is easy to design it might be difficult to granularize the tagset for various requirements, which might require different degree of fineness. Most of the popular English tagsets such as Upenn, Brown, C5 are all flat tagsets, which could be reflective of the simpler morphosyntatctic nature of the language.

Hierarchical tagset on the other hand offers an elegant way to design the tagset and has several distinct layers, each additional layer adding another level of granularity. It would be desirable to have a discussion on the type of tagset appropriate for Indic languages given the nature of the Indian languages such as richer morphology, agglutination etc. and the availability of basic tools in these languages.

We have to keep in mind that, the new tagset should be mappable to the existing tagsets mentioned above. The mapping should be at least unidirectional if not bidirectional. This would

probably mean manually creating the mapping table and identifying rules for ambiguous cases, if required.

2.4. CIIL HINDI CORPORA TAGSET FINALIZED ON AUGUST 29, 2006

	Description	**TAG**
NOUN		
1	Common Noun	NN
2	Proper Noun (place, person, etc.)	NNP
3	Noun compound	NC
4	Noun Abstract	NAB
5	Cardinal number	CRD
6	Ordinal number	ORD
PRONOUN		
7	Personal Pronoun	PRP
8	Indefinite Pronoun	PRI
9	Pronoun Reflexive	PRR
10	Relative Pronoun	PRL
11	Demonstrarive Pronoun	PDP
	VERB	

12	Verb Finite Main	VF
13	Verb non-finite Adjectival Verb non-finite Adverbial	VNF
14	Verb Auxiliary	VAX
15	Gerund/Verb non-finite Nominal	VNN
16	Verb infinitive	VINF
17	Verb Causative	VCC
18	Verb Double Causative	VCD
	ADJECIVE	
19	Adjective - declinable	ADD
20	Adjective – indeclinable	ADI
ADVERB		
21	Adverb	ADV
22	Adverb of location	ADL
		OTHERS
23	Question Words	QWO
24	Post Position	PPO
25	Intensifier	INT
26	Quantifier	QTF

27	Conjunction	CNJ
28	Interjection	INJ
29	Indeclinable	IND
30	Quotative	QOT
31	Particle	PRT
32	Negative	NEG
33	Reduplicative	RDP
34	Loan Word (non-nativized words)	FWD
35	Idiom	IDM
36	Proverb	PRO

2.5. AU-KBC POST TAGSET FOR TAMIL

NOUN	Description	TAG
1	Noun Singular	N.Sg
2	Noun Plural	N.Pl
3	Ponoun	Pr.Sg
4	Pronoun Oblique Form	Pr.Obl.Sg
5	Pronoun Plural	Pro.Pl

6	Pronoun Plural Oblique	Pro.Obl.Pl
7	#Euphonic Increment	Eu
8	Accusative	Acc
9	Dative	Dat
10	Instrumental	Inst
11	Sociative	Soc
12	Locative	Loc
13	Ablative	Abl
14	Benefactive	Ben
15	Genitive	Gen
16	Vocative	Voc
17	Clitics	Cl
VERB		
18	Verb	V
19	#Transitive Verb	VT
20	#Intransitive Verb	VI
21	#Causative	VC
22	Infinitive	Inf
23	Singular Negative (imperative)	Neg
24	Plural Imperative	V.Pl

25	Plural Negative (imperative)	Pl.Neg
26	Imperative + Suffixes	Imp
27	Present Tense	Pre
28	Past Tense	Pst
29	Future Tense	Fut
30	First Person Singular	1S
31	First Person Plural	1Pl
32	Second Person Singular	2S
33	Second Person Plural/Honorific	2Pl
34	Third Person Masculine Singular	3MS
35	Third Person Feninine Singular	3FS
36	Third Person Honorific Singular	3HS
37	Third Person Honorific Plural	3HPl
38	Third Person Singular Neuter (non-human)	3NS
39	Third Person Plural Neuter (non-human)	3NPl
40	Third Person Neuter Future Negative	3NFutNeg
41	Future Negative Human	HumFutNeg
42	Present/Past Negative (non-future)	NfutNeg
43	#Optative	Opt

44	Verbal Participle	Vbp
45	Verbal Paticiple Negative	VbpNeg
46	Conditional Positive	Cnd
47	Conditional Negative	Cnd Neg
48	Adjectival Participle Present/Past	AdjP
49	Adjectival Participle Future	AdjPP
50	Adjectival Negative	AdjPNeg
51	Verbal Noun Untensed	VbnUnt
52	Verbal Noun Tensed Positive	Vbn
53	Verbal Noun Tensed Negative	VbnNeg
54	Paticipial Noun	PplN
55	Adjectival Noun	AdjN
	Selective (Concessive?)	Sel
57	Ordinal Suffix	Ord
58	Reportative	Rep
59	Complementizer	Cmp
60	Auxiliary	Aux
61	Conjunction	Cnj
62	Adjectives	Adj
63	Adverbs	Adv

64	Postpositions	PSP
65	Quantifiers	Q
66	Determiners	Det
67	Wh Words	INT
68	Unkonwn	UNK

2.6. IIIT-H

	Description	TAGs
NOUN		
1	Common Noun	NN
2	Proper Noun	NNP
3	Cardinal Number	QC
4	Ordinal Number	QO
5	Locative Noun	NST
PRONOUN		
6	Personal Pronoun	PRP
VERB		
7	Finite Verb	VM
8	Verb Auxiliary	VAUX
ADJECTIVE		

9	Adjective	JJ
ADVERB		
10	Adverb	RB
OTHERS		
11	Question Words	WQ
12	Postposition	PSP
13	Intensifier	INTF
14	Quantifier	QF
15	Conjuction	CC
16	Interjection	INJ
17	Quotative	UT
18	Particle	RP
19	Negative	NEG
20	Reduplication	RDP
21	Foreign Words	UNK
22	Symbols	SYM
23	Echo Words	ECH
24	Compounds	*C
25	Demonstrative	DEM
26	Classifier	CL

Amrita Tagset

S.N	TAG	DESCRIPTION	S.N	TAG	DESCRIPTION
1	<NN>	NOUN	17	<VINT>	VERB INFINITE
2	<NNC>	COMPOUND NOUN	18	<CNJ>	CONJUNCTION
3	<NNP>	PROPER NOUN	19	<CVB>	CONDITIONAL VERB
4	<NNPC>	COMPOUND PROPER NOUN	20	<QW>	QUESTION WORDS
5	<ORD>	ORDINALS	21	<COM>	COMPLEMENTIZER
6	<CRD>	CARDINALS	22	<NNQ>	QUANTITY NOUN
7	<PRP>	PERSONAL PRONOUN	23	<QTF>	QUANTIFIERS
8	<PRIN>	PRONOUN INTROGATIVE	24	<PPO>	POSTPOSITIONS
9	<PRID>	PRONOUN INDEFINITE	25	<DET>	DETERMINERS
10	<ADJ>	ADJECTIVE	26	<INT>	INTENSIFIER
11	<ADV>	ADVERB	27	<ECH>	ECHO WORDS
12	<VNAJ>	VERB NON FINITE ADJECTIVE	28	<EMP>	EMPHASIS
13	<VNAV>	VERB NON FINITE ADVERB	29	<COMM>	COMMA
14	<VBG>	VERBAL GERUND	30	<DOT>	DOT
15	<VF>	VERB FINITE	31	<QM>	QUESTION MARK
16	<VAX>	VERB AUXILARY	32	<RDW>	REDUPLICATION WORDS

2.7. Comparison

Nouns

AU-K BC	**CIIL**	**IIIT-H**
	NN (Common Noun)	NN (Common Noun)
	NNP (Proper Noun)	NNP (Proper Noun)
	NC (Noun Compound)	*C (For all compounds)
Ord (Ordinal suffix)	CRD (Cardinal No.)	QC (Cordinal No.)
	ORD (Ordinal No.)	QO (Ordinal No.)
N.Sg (Noun Sg)		

N.Pl (Noun Plural)		
	PRP (Personal Pronoun)	PRP (Pronoun)
	PRI (Indefinite Pronoun)	
	PRR (Reflexive Pronoun)	
	PDP (Demonstrative)	
Pr.Sg. (Pronoun Sg)		
Pr.Pl. (Pronou Plural)		
Pr.Obl.Sg. (Pronoun Oblique Sg.)		
Pr.Obl.Pl (Pronoun Oblique Plural)		

Verbs

AU-KBC	**CILL**	**IIIT-H**
	VF (Verb Finite Main)	VF (Verb Finite Main)
	VNF (Verb Non-finite Adverbial and adjectival	VNF (Verb Non-Finite adverbial and adjectival)
	Vax (Verb Auxiliary)	VAUX (Verb Auxiliary
VbnUnt (Verbal non Untensed) Vbn (Vgb Tensed Positive	VNN (Gerund/Verb non-finite nominal)	VNN (Gerund/Verb non-finite nominal)

VbnNeg (Vbn Tensed Negative)		
Inf (Infinitive)	VINF (Verb Infinitive)	VINF (Verb Infinitive)
VC (Verb Causative)	VCC (Verb Causative)	VINF (Verb Infinitive)
	VCD (Verb Double Causative)	
VT (Verb Transtitive)		
VI (Verb Intransitive)		

Adjective

AU-KBC	**CILL**	**IIIT-H**
Adj (Adjectives)		JJ (Adjectives)
	ADD (Adjective Declinable)	
	ADI (Adjective Indeclinable)	

Adverbs

AU-KBC	**CILL**	**IIIT-H**
Adv (Adverbs)	ADV (Adverbs)	RB (Adverbs)
	ADL (Adverb of location)	

Others

AU-KBC	**CILL**	**IIIT-H**
PSP (Postposition)	PRO (Postposition)	PSP (Postposition)
Q (Quantifiers)	QTF (Quantifiers)	QF (Quantifiers)
INT (Wh-words)	QWO (Question words)	WQ (Qestion words)
Cnj	CNJ (Conjunction)	CC (conjunct)
	INT (Intensifier)	INT (Intensifier)
	NEG (Negative)	NEG (Negative)
	INJ (Interjection)	INJ (Interjection)
	PRT (Particle)	RP (Particle)
	IND (Indeclinable)	
	QOT (Quotative)	
	RDP (Reduplication)	
	FWD (Loan Word)	
	IDM (Idiom)	
	PRO (Proverb)	
		CL (Classifier)
		SYM (Special)

2.8. Tamil Tagset (full framework)

Level-0	**Level-1**	**Level 2**

Lexical categories	Types	Attributes	Values	Types Applicable	Remarks
[N] Nouns	[1] Common [2] Proper [3] Verbal [4] Spatial/ temporal Adjectival nouns such as **நல்லவன்** will be tagged as common nouns	Gender	[1] Masculine [2] Feminine [3] Neuter	[1] Common	
		Number	[1] Singular [2] Plural [3] Plural/ honourific	[1] Common [2] Proper	
		Case*	[1] Direct [2] Oblique	[1] Common	
		Case-markers	[2] Accusative (*E*) [3] Instrumental (*Al*) [4] Dative (*kku* or *ukku*) [5] Genitive (*in* | *utEya* | *athu*) [6] Sociative (*otu* | *utan*) [7] Locative (*il* | *itam*) [8] Ablative (*irunthu*) [9] Benefactive (*ukkAka*)	[1] Common [2] Proper [3] Verbal [4] Spatial/ termporal	

		Negative	[1] Yes	[1] Common [2] Proper [4] Spatial/ termporal	
		Emphatic*	[1] Yes	[1] Common [2] Proper [4] Spatial/ termporal	
		Definiteness*	[1] Yes	[1] Common [2] Proper [4] Spatial/ termporal	
[V] Verbs	[1] Finite [2] Auxiliary [3] Infinitive [4] Nonfinite Non-finites forms will betagged as participles	Gender	[1] Masculine [2] Feminine [3] Neuter	[1] Finite [2] Auxiliary	
		Number	[1] Singular [2] Plural [3] Plural/ honourific	[1] Finite [2] Auxiliary	
		Person	[1] First [2] Second [3] Third	[1] Finite [2] Auxiliary	
		Tense	[1] Past [2] Present [3] Future	[1] Finite [2] Auxiliary	

		Aspect*	[1] Simple [2] Continuous (progressive) [3] Perfect	[2] Auxiliary	Dravidian has separate auxiliaries for aspect and mood. However, many a times these are agglutinated with the main verb. Here it is recommended that the word be segmented first before being tagged.
		Mood*	[1] Declarative [2] Subjunctive [3] Conditional [4] Imperative [5] Presumptive	[2] Auxiliary	
		Negative	[1] Yes	[1] Finite [2] Auxiliary [3] Nonfinite	
		Emphatic*	[1] Yes	[1] Finite [2] Auxiliary [3] Infinitive	
[PR] Pronouns	[1] Personal [2] Possessive [3] Reflexives	Gender	[1] Masculine [2] Feminine [3] Neuter	[1] Personal [2] Possessive	

	[4] Reciprocals [5] Spatial/ temporal	Number	[1] Singular [2] Plural [3] Plural/honourific	[1] Personal [2] Possessive	
		Person	[1] First [2] Second [3] Third	[1] Personal [2] Possessive	
		Case	[1] Direct [2] Oblique	[1] Personal [2] Possessive	
		Case-markers	[2] Accusative (*E*) [3] Instrumental (*Al*) [4] Dative (*kku* or *ukku*) [5] Genitive (*in* | *utEya* | *athu*) [6] Sociative (*otu* | *utan*) [7] Locative (*il* | *itam*) [8] Ablative (*irunthu*) [9] Benefactive (*ukkAka*)	[1] Personal [2] Possessive [3] Reflexive [5] Spatial/ temporal	
		Referential	[1] Absolute [2] Relative [3] Wh-words	[1] Personal [2] Possessive	

		Negative	[1] Yes	[1] Personal [2] Possessive [5] Spatial/ temporal	
		Emphatic*	[1] Yes	[1] Personal [2] Possessive [5] Spatial/ temporal	
		Definiteness*	[1] Yes	[1] Personal [2] Possessive [5] Spatial/ temporal	
		Incl/ Excl.*	[1] Yes	[1] Personal	
[NM] Nominal modifiers	[1] Adjectives [2] Demonstratives [3] Quantifiers	Referential	[1] Absolute [2] Relative [3] Wh-words	[2] Demonstratives	
		Numeral	[1] Non-numeral [2] Cardinal number [3] Ordinal number	[3] Quantifier	
		Negative	[1] Yes	[1] Adjectives [3] Quantifiers	
		Emphatic*	[1] Yes	[2] Demonstratives [3] Quantifiers	

[RB] Adverbs	-	Negative	[1] Yes	-	
[PL] Participles	[1] Relative [2] Verbal [3] Nominal [4] Conditional	Gender	[1] Masculine [2] Feminine [3] Neuter	[3] Nominal	*ceykiRavan, ceythathu, ceythavarkaL, ceypavar* Should these words be tagged as nominal participles?
		Number	[1] Singular [2] Plural [3] Plural/ honourific	[3] Nominal	
		Case-markers	[2] Accusative (*E*) [3] Instrumental (*Al*) [4] Dative (*kku* or *ukku*) [5] Genitive (*in* | *utEya* | *athu*) [6] Sociative (*otu* | *utan*) [7] Locative (*il* | *itam*) [8] Ablative (*irunthu*) [9] Benefactive (*ukkAka*)	[3] Nominal	

		Tense	[1] Past [2] Present [3] Future	[1] Relative [3] Nominal	
		Aspect*	[1] Simple [2] Continuous [3] Perfect	[3] Nominal	
		Negative	[1] Yes	[2] Relative [3] Nominal [4] Conditional	
		Emphatic*	[1] Yes	[3] Nominal [4] Conditional	
		Realis*	[1] Realis [2] Irrealis	[4] Conditional	Marks if the event is real or hypothesised one. Eg. **வந்ததால்** *vs.* **வந்தால்**
[PP] Postpositions	-	Emphatic*	[1] Yes	-	

[RP] Particles	[1] Coordinating [2] Subordinating [3] Interjection [4] Others	-	-	-	
[PU] Punctuation	-	-	-	-	
[RD] Residual	-	-	-	-	

* Optional attributes

2.9. Comparison of Tagset for Indian Languages

A comparative study of the CIIL-KHS Hindi Corpora Tag set, IIIT Tag set and the Penn Treebank Tag sets are performed. The three tag sets have many common tags.

		CIIL-KHS	**PENN**	**IIIT**
NOUNS	Common Noun	NN	NN	NN
	Proper Noun (Place, person)	NNP	NNP-Sg, NNPS-Pl	NNP
	Compound Noun (aCHanammamaar)	NC		
	Abstract Noun (maNam)	NAB		
	Cardinal Number	CRD	CD	QC
	Ordinal Number	ORD		QO
	Plural Noun		NNS	

	Possesive Noun		POS	
	Locative Noun			NST
PRONOUNS	Personal Pronoun	PRP	PRP	PRP
	Indefinite Pronoun	PRI		
	Pronoun Reflexive	PRR		
	Relative Pronoun	PRL		
	Demonstrative Pronoun	PDP		
	Possesive Pronoun		PRP$	
VERBS	Finite Verb	VF	VB	VM
	Verb Auxiliary	VAX		VAUX
	Verb Nonfinite	VNF		
	Verb Infinite	VINF		
	Verb Gerund(present participle)	VNN	VBG	
	Verb Causative	VCC		
	Verb Double Causative	VCD		
	Verb, Past tense		VBD	

	Verb, Past participle		VBN	
	Verb,non-3rd person sing.present		VBP	
	Verb,3rd person sing.present		VBZ	
ADJECTIVES	Adjective-declinable	ADD	JJ	JJ
	Adjective-indeclinable	ADI		
	Adjective-comparative		JJR	
	Adjective-superlative		JJS	
ADVERBS	Adverb	ADV	RB	RB
	Adverb of Location	ADL		
	Adverb,Comparitive		RBR	
	Adverb, Superlative		RBS	
OTHERS	Question Words	QWO		WQ
	Post Position	PPO		PSP
	Intensifier	INT		INTF

	Quantifier	QTF		QF
	Conjunction	CNJ	CC	CC
	Interjection	INJ	UH	INJ
	Quotative	QOT		UT
	Particles	PRT	RP	RP
	Negative	NEG		NEG
	Reduplication	RDP		RDP
	Foreign Words	FWD	FW	UNK
	Symbols		SYM	SYM
	Echo Words			ECH
	Compounds			*C
	Wh-determiner(which,that)		WDT	
	Wh-pronoun(what,who,whom)		WP	
	Possesive wh-pronoun(whose)		WP$	
	Wh-adverb(how,where,why)		WRB	
	Idiom	IDM		
	Proverb	PRO		
	Indeclinable	IND		

	To		TO	
	Determiner		DT	
	Existential There(There is)		EX	
	Preposition or Subordinating conjunction		IN(in, of like)	
	List Item Marker		LS	
	Modal		MD	
	Predeterminer		PDT	
	Demonstrative			DEM
	Classifier			CL

2.3.2. Discussion is Tag design

1. Introduction
2. Objective
3. Some Assumptions
4. Issues in Tag Set Design

    4.1 Fineness vs Coarseness

    4.2 Syntactic Function vs Lexical Category

    4.3 New Tags vs Tags from a Standard Tagger

5. POS Tags Chosen for the Current Scheme

6. Some Special Cases
   6.1 'vAlA' type constructions
   6.2 Honorifics in Indian languages
   6.3 Foreign words
7. A Special Note
8. Chunk Tags Chosen for the Current Scheme
   8.1 Definition of a chunk
   8.2 Chunk Types
   8.3 Some Special Cases
9. Annotation method/procedure
10. Conclusion
11. References
12. Acknowledgment
13. Appendix
   13.1 List of POS Tags
   13.2 List of Chunk Tags

2.3.2.1. Introduction

The significance of large annotated corpora in the present day NLP is widely known. Annotated corpora serve as an important tool for investigators of natural language processing, speech recognition and other related areas. It proves to be a basic building block for constructing statistical models for automatic processing of natural languages.

Many such corpora are available for languages across the world and have proved to be a useful step towards natural language processing. Coming to the scenario for Indian languages, not much work has been carried out on the front of automatic processing of Hindi or any other Indian language. The main bottleneck being unavailability of an annotated corpora, large enough to experiment statistical algorithms.

Annotation of corpora (AnnCorra) can be done at various levels viz, part of speech, phrase/clause level, dependency level, etc. Part of speech tagging forms the basic step towards building an annotated corpus. Chunking can form the next level of tagging.

The task of annotating corpora of several Indian languages has been taken up in the Indian Language Machine Translation (ILMT) project. ILMT is a project in which a number of institutes have come together to form a consortium and work towards developing MT systems for various Indian language pairs.

A primary requirement of such an effort is defining standards for various sub tasks. Thus, standardization of annotation schemes for various annotation tasks becomes a crucial step in this direction.

The issues related to defining standards for POS/Chunk tagging schemes were discussed by scholars from various Indian institutes by way of holding meetings etc. and some standards have been arrived at.

2.3.2.2. Objective

The purpose of the workshop was to arrive at standard tagging scheme for POS tagging and chunking for annotating Indian languages (AnnCorraTS) and come up with the tags which are exhaustive for the task of annotation for a larger group of languages, specially, Indian languages. The present document gives a detailed description of the tags which have been defined for the tagging schemes and elaborates the motivations behind the selection of these tags. The document also discusses various issues that were addressed while preparing the tag sets and how they have been resolved.

2.3.2.3. Some Assumptions

During the workshop it was decided to base the discussion and decisions about various tags on the following basic assumptions which everybody agreed on:

i) The tags should be common for all Indian languages

ii) It should be comprehensive/ complete

iii) It should be simple. Maintaining simplicity is important for the following two reasons :

(a) Ease of Learning

(b) Consistency in annotation

Another important point which was discussed and agreed upon was that POS tagging is NOT a replacement for morph analyser. A 'word' in a text carries the following linguistic knowledge

a) grammatical category and

b) grammatical features such as gender, number, person etc. The POS tag should be based on the 'category' of the word and the features can be acquired from the morph analyser.

2.3.2.4. Issues in Tag Set Design

This section deals with some of the issues related to any POS tagger and the policy that we have adopted to deal with each of these issues for our purpose.

The first step towards developing POS annotated corpus is to come up with an appropriate tags. The major issues that need to be resolved at this stage are :

1. Fineness vs Coarseness in linguistic analysis
2. Syntactic Function vs lexical category
3. New tags vs tags from a standard tagger

2.3.2.4.1 Fineness vs Coarseness

An issue which always comes up while deciding tags for the annotation task is whether the tags should capture 'fine grained' linguistic knowledge or keep it 'coarse'. In other words, a decision has to be taken whether or not the tags will account for finer distinctions of the parts of speech features. For example, it has to be decided if plurality, gender and other such information will be marked distinctly or only the lexical category of a given word should be marked.

It was decided to come up with a set of tags which avoids 'finer' distinctions. The motivation behind this is to have less number of tags since less number of tags lead to efficient machine learning. Further, accuracy of manual tagging is higher when the number of tags is less.

However, an issue of general concern is that in an effort to reduce the number of tags we should not miss out on crucial information related to grammatical and other relevant linguistic knowledge which is encoded in a word, particularly in agglutinating languages, eg, Tamil, Telugu and many other Indian languages. If tags are too coarse, some crucial information for further processing might be missed out. As mentioned above, primarily the required knowledge for a given lexical item is its grammatical category, the features specifying its grammatical information and any other information suffixed into it. For example,

Telugu word ' *rAmudA* (Is it Ram ?)' contains the following information <category (noun)+grammatical features(masculine, singular) + question>. The word by itself is a bundle of linguistic information. Morph analyser provides all the knowledge that is contained in a word. It was decided that any linguistic knowledge that can be acquired from any other source (such as morph analyser) need not be incorporated in the POS. As mentioned above, POS tagger is not a replacement for morph analyser. In fact, features from morph analyser can be used for enhancing the performance of a POS tagger. The additional knowledge of a POS given by a POS tagger can be used to disambiguate the multiple answers provided by a morph analyser.

On the other hand, we agree that too coarse an analysis is not of much use. Essentially, we need to strike a balance between fineness and coarseness. The analysis should not be so fine as to hamper machine learning and also should not be so coarse as to miss out important information. It is also felt that fine distinctions are not relevant for many of the applications(like sentence level parsing, dependency marking, etc.) for which the tagger may be used in future.

However, it is well understood that plurality and other such information is crucial if the POS tagged corpora is used for any application which needs the agreement information. In case such information is needed at a later stage, the same tag set can be extended to encompass information

such as plurality etc as well. This can be done by providing certain heuristics or linguistic rules.

Thus, to begin with, it has been decided to adopt a coarse part of speech analysis. At the same time, wherever it is found essential, finer analysis is incorporated. Also, there is a basic understanding that wherever/whenever essential, the tags containing finer linguistic knowledge can be incorporated. An example of where finer analysis becomes crucial has been given below. Take the Hindi sentence (h1) below :

h1. *AsamAna*_NN *meM*_PSP *uDane*_VM *vAlA*_PSP *ghoDA*_NN
'sky' 'in' 'flying' 'horse
*nIce*_NST *utara*_VM *AyA*_VAUX.
'down' "descend" "came"

In (h1) above *uDane* is a noun derived from a verb. The word *AsamAna* is an argument of *uDane* and not of *'nIce utara AyA* – another verb in the sentence. It is crucial to retain the information that *uDane,* though functioning as a noun now, is derived from a verb and can take its own arguments. In order to preserve such crucial information a finer analysis is essential. Therefore, a distinct tag needs to be introduced for such expressions. In the current tagging scheme *uDane* will be annotated as a 'main verb (VM)' at the POS level. However, the information that it is functioning like a noun will be captured at the chunking level by introducing a distinct chunk tag VGNN (discussed in details under Section III on Chunking).

2.3.2.4.2 Syntactic Function vs Lexical Category

A word belonging to a particular lexical category may function differently in a given context. For example, the lexical category of *harijana* in Hindi is a noun . However, functionally, *harijana* is used as an adjective in (h2) below,

h2. *eka dina pAzca baje khabara AyI ki koI* ***harijana***
'one' 'day' 'five' 'o'clock' 'mews' 'came' 'that' 'some' 'harijana'
bAlaka *unase milanA cAhatA hE*

'young boy' 'him' 'to meet' 'wants' 'is'

“One day, a message came at five o'clock that some 'harijana' boy wanted to meet him”.

Such cases require a decision on whether to tag the concerned word according to its lexical category or by its syntactic category. Since the word in a context has syntactic relevance, it appears natural to tag it based on its syntactic information. However, such a decision may lead to further complications.

In AnnCorra, the syntactic function of a word is not considered for POS tagging. Since the word is always tagged according to its lexical category there is consistency in tagging. This reduces confusion involved in manual tagging. Also the machine is able to establish a word-tag relation which leads to efficient machine learning.

In short, it was decided that syntactic and semantic/pragmatic functions were not to be the basis of deciding a POS tag.

2.3.2.4.3. New Tags vs Tags from a Standard Tagger

Another point that was considered while deciding the tags was whether to come up with a totally new tag set or take any other standard tagger as a reference and make modifications in it according to the objective of the new tagger. It was felt that the later option is often better because the tag names which are assigned by an existing tagger may be familiar to the users and thus can be easier to adopt for a new language rather than a totally new one. It saves time in getting familiar to the new tags and then work on it.

The Penn tags are most commonly used tags for English. Many tag sets designed subsequently have been a variant of this tag set (eg. Lancaster tag set). So, while deciding the tags for this tagger, the Penn tags has been used as a benchmark. Since the Penn tag set is an established tag set for English, we have used the same tags as the Penn tags for common lexical types. However, new tags have been introduced wherever Penn tags have been found inadequate for Indian

language descriptions. For example, for verbs none of the Penn tags have been used. Instead AnnCorra has only two tags for annotating verbs, VM (main verb) and VAUX (auxiliary verb).

2. **3.2.5. POS tags Chosen for the Current Scheme**

This section gives the rationale behind each tag that has been chosen in this tag set.

2.3.2.5.1 NN Noun

The tag NN for nouns has been adopted from Penn tags as such. The Penn tag set makes a distinction between noun singular (NN) and noun plural (NNS). As mentioned earlier, distinct tags based on grammatical information are avoided in IL tagging scheme. Any information that can be obtained from any other source is not incorporated in the POS tag. Plurality, for example, can be obtained from a morph analyzer. Moreover, as mentioned earlier, if a particular information is considered crucial at the POS tagging level itself, it can be incorporated at a later date with the help of heuristics and linguistic rules. This approach brings the number of tags down, and helps achieve simplicity, consistency, better machine learning with a small corpora etc. Therefore, the current scheme has only one tag (NN) for common nouns without getting into any distinction based on the grammatical information contained in a given noun word.

2.3.2.5.2 NST Noun denoting spatial and temporal expressions

A tag NST has been included to cover an important phenomenon of Indian languages. Certain expressions such as '*Upara*' (above/up), '*nIce*' (below) '*pahale*' (before), 'Age' (front) etc are content words denoting time and space. These expressions, however, are used in various ways. For example,

5.2.1 These words often occur as temporal or spatial arguments of a verb in a given sentence taking the appropriate *vibhakti* (case marker):

h3. *vaha **Upara** so rahA thA .*
'he' 'upstairs' 'sleep' 'PROG' 'PAST'
"He was sleepign upstairs".

h4. *vaha **pahale** **se** kamare meM bEThA thA .*
'he' 'beforehand' 'from' ' room' 'in' 'sitting' 'was'
"He was sitting in the room from beforehand"

h5. *tuma **bAhara** bETho*
'you' 'outside' 'sit'
"You sit outside".

Apart from functioning like an argument of a verb, these elements also modify another noun taking postposition 'kA'.

h6. *usakA baDZA bhAI **Upara** **ke** **hisse** **meM** rahatA hE*
'his' 'elder' 'brother' 'upstairs' 'of' 'portion' 'in' 'live' 'PRES'
"His elder brother lives in the upper portion of the house".

5.2.2 Apart from occuring as a nominal expression, they also occur as a part of a postposition along with 'ke'. For example,

h6. *ghaDZe **ke** **Upara** thAlI rakhI hE.*
'pot' 'of' 'above' 'plate' 'kept' 'is'
The plate is kept on the pot".

h7. *tuma ghara ke **bAhara** bETho*
'you' 'home' 'of' 'outside' 'sit'

"You sit outside the house".

'Upara' and '*bAhara*' are parts of complex postpositions *'ke Upara'* and '*ke bAhara*' in (h6) and (h7) respectively which can be translated into English prepositions 'on' and 'outside'.

For tagging such words, one possible option is to tag them according to their syntactic function in the given context. For example in 5.2.2 (h6) above, the word '*Upara*' is occurring as part of a postposition or a relation marker. It can, therefore, be marked as a postposition. Similarly, in 5.2.1. (h3-4) and (h6) above, it is a noun, therefore, mark it as a noun and so on. Alternatively, since these words are more like nouns, as is evident from 5.2.1 above they can be tagged as nouns in all there occurrences.

However, if we follow any of the above approaches we miss out on the fact that this class of words is slightly different from other nouns. These are nouns which indicate 'location' or 'time'. At the same time, they also function as postpositions in certain contexts. Moreover, such words, if tagged according to their syntactic function, will hamper machine learning. Considering their special status, it was considered whether to introduce a new tag, NST, for such expressions. Following five possibilities were discussed :

a) Tag both (h7) & (h8) as NN

b) Tag both (h7) & (h8) as NST

c) Tag (h7) as NN & (h8) as NST

d) Tag (h7) as NST & (h8) as PSP

e) Tag (h7) as NN & (h8) as PSP

After considering all the above, the decision was taken in favour of (b). The decision was primarily based on the following observations:

(i) *'bAhara'* in both (h7) and (h8) denotes the same expression (place expression 'outside')
(ii) In both (h7) and (h8), *'bAhara'* can take a vibhakti like a noun (**bAhara ko** bETho, ghara **ke bAhara ko** bETho)
(iii) If a single tag is kept for both the usages, the decision making for annotators would also be easier.
Therefore, a new tag **NST** is introduced for such expressions. The tag **NST** will be used for a finite set of such words in any language. For example, Hindi has

Age (front), *pIche* (behind), *Upara* (above/upstairs), *nIce* (below/down), *bAda* (after), *pahale* (before), *andara* (inside), *bAhara (outside)* etc.

2.3.2.5.2 NNP Proper Nouns

The need for a separate tag for proper nouns and its usability was discussed. Following points were raised against the inclusion of a separate tag for proper nouns :

a) Indian languages, unlike English, do not have any specific marker for proper nouns in orthographic conventions. English proper nouns begin with a capital letter which distinguishes them from common nouns.
b) All the words which occurs as proper nouns in Indian languages can also occur as common nouns denoting a lexical meaning. For example,
English : John, Harry, Mary occur only as proper nouns whereas
Hindi : *aTala bihArI, saritA, aravinda* etc are used as 'names' and they also belong to grammatical categories of words with various senses . For example given below is a list of Hindi words with their grammatical class and sense.

aTala	adj	immovable
bihArI	adj	from Bihar
saritA	noun	river
aravinda	noun	lotus

Any of the above words can occur in texts as common lexical items or as proper names. (h9) - (h11) below show their occurrences as proper nouns,

h9. ***atala bihAri bAjapaI*** *bhArata ke pradhAna mantrI the*.
'Atal' 'Bihari' 'Vajpayee' 'India' 'of' 'prime' 'minister' 'was'
"Atal Behari Vajpayee was the Prime Minister of India".

h10. merI mitra **saritA** tAIvAna jA rahI hE.
'my' 'friend' 'Sarita' 'Taiwan' 'go' 'PROG' 'is'
"My friend Sarita is going to Taiwan"

h11. ***aravinda*** *ne mohana ko kitAba dI*.
'Aravind' 'erg' 'Mohan' 'to' 'book' 'gave'
"Aravind gave the book to Mohan".

Therefore, in the Indian languages' context, annotating proper nouns with a separate tag will not be very fruitful from machine learning point of view. In fact, the identification of proper nouns can be better achieved by named entity filters.

Another point that was considered in this context was the effort involved in manual tagging of proper nouns in a given text. It is felt that not much extra effort is required in manual tagging of proper nouns. However, the data annotated with proper nouns can be useful for certain applications. Therefore, there is no harm in marking the information if it does not require much effort.

Finally, it was decided to have a separate tag for proper nouns for manual annotation and ignore it for machine learning algorithms. Following this decision, the tag **NNP** is included in the tag set. This tag is the same as the Penn tag for proper nouns. However, in this case also AnnCorraTS has only one tag for both singular and plural proper nouns unlike Penn tags where a distinction is made between proper noun singular and proper noun plural by having two tags

NNP and NNPS respectively.

2.3.2.5.3.1. PRP Pronoun

Penn tags make a distinction between personal pronouns and possessive pronouns. This distinction is avoided here. All pronouns are marked as PRP. In Indian languages all pronouns inflect for all cases (accusative, dative, possessive etc.). In case we have a separate tag for possessive pronouns, new tags will have to be designed for all the other cases as well. This will increase the number of tags which is unnecessary. So only one tag is used for all the pronouns. The necessity for keeping a separate tag for pronouns was also discussed, as linguistically, a pronoun is a variable and functionally it is a noun. However, it was decided that the tag for pronouns will be helpful for anaphora resolution tasks and should be retained.

2.2.5.3.2. DEM Demonstratives

The tag 'DEM' has been included to mark demonstratives. The necessity of including a tag for demonstratives was felt to cover the distinction between a pronoun and a demonstrative. For example,

h12. ***vaha ladakA*** *merA bhAI hE* (demnostrative)
'that' 'boy' 'my' 'brother' 'is'
h13. ***vaha*** *merA bhAI hE*(pronoun)
'he' 'my' 'brother' 'is'

Many Indian languages have different words for demonstrative adjectives and pronouns. A better evidence for including a separate tag for demonstratives is from the following Telugu examples,

t1. *A abbAyi nA tammudu*
'that' 'boy' 'my' 'brother'

t2. *atanu nA tammudu*

'he' 'my' 'brother'

(Telugu does not have a copula 'be' in the present tense)

2.2.5.4. VM Verb Main

Verbal constructions in languages may be composed of more than one word sequences. Typically, a verb group sequence contains a main verb and one more auxiliaries (V AUX AUX ... ...). In the current tagging scheme the support verbs (such as *dAlanA* in *kara dAlAtA hE*, *uThanA* in *cOMka uThA thA* etc) are also tagged as VAUX. The group can be finite or non-finite. The main verb need not be marked for finiteness. Normally, one of the auxiliaries carries the finiteness feature.

The necessity of marking the finiteness or non-finiteness in a verb was discussed extensively and everybody agreed that it was crucial to mark the distinction. However, languages such as Hindi, which have auxiliaries for marking tense, aspect and modalities pose a problem. The finiteness of a verbal expression is known only when we reach the last auxiliary of a verb group. Main verb of a finite verb group (leaving out the single word verbal expressions of the finite type – eg *vaha dillI gayA*) does not contain finiteness information. For example,

h14. *laDZakA seba* ***khAtA*** ***raHA wA***

'boy' 'apple' 'eating' 'PROG' 'was'

The boy had kept eating.

h15. *seba* ***khAtA*** ***huA*** *laDZakA jA rahA thA*

'apple' 'eating' 'PROG' 'boy' ' go' 'PROG' 'was'

The boy eating the apple was going.

The expression *khAtA raHA* in (h29) above is finite and *khAtA huA* in (h3) is non finite. However, the main verb *'khAwA*' is non-finite in both the cases.

So, the issue is - whether to (1a) mark finiteness in "***khAwA rahA wA (*** had kept eating***)***" at the lexical level on the main verb (khA) or (1b) on the auxiliary containing finiteness (wA) or (2) not mark it at the lexical level at all. All the three possibilities were discussed;

1) Mark the finiteness at the lexical level.

If we mark it at the lexical level, following possibilities are available :

1a) Mark the finiteness on the main verb, even though we know that the lexical item itself is not finite.

In this case, the annotator interprets the finiteness from the context. (The POS tags VF, VNF and VNN were earlier decided based on this approach). The main verb, therefore, is marked as finite consciously with a view that the group contains a 'verb root' and its auxiliaries (as TAM etc) and although the main verb does not carry the finiteness at the lexical level, the verb group is finite. Although this approach facilitates annotation of both, the main verb and the finiteness (of the group) by a single tag, it allows tagging a lexical item (main verb) with the finiteness feature which it does not actually carry. So, this is not a neat solution.

1b) The second possibility is, mark the finiteness on the last auxiliary of the sequence. Here again the decision has to be taken from the context. This possibility was not considered since this also involves marking the verb finiteness at the lexical level.

2) Don't mark the finiteness at the lexical level. Instead mark it as indicated in (2a) or (2b) below.

2a) Introduce a new layer which groups the verb group and mark the verb group as finite or non-finite. This approach proposes the following :

(i) Annotate the main verb as **VM** (introduce a new tag). Thus,

h14a. *laDZakA seba* ***khAtA*_VM** *raHA wA*
'boy' 'apple' 'eating' 'PROG' 'was'

h15a. *seba* ***khAtA*_VM** *huA laDZakA jA rahA thA*
'apple' 'eating' 'PROG' 'boy' ' go' 'PROG' 'was'

(ii) Annotate the auxiliaries as **VAUX**,

h14a. *laDZakA seba khAtA*_VM *raHA*_**VAUX** *wA*_**VAUX**
'boy' 'apple' 'eating' 'PROG' 'was'
h15a. *seba khAtA*_VM *huA*_**VAUX** *laDZakA jA rahA thA*
'apple' 'eating' ' PROG' 'boy' ' go' 'PROG' 'was'

(iii) Group the verb group (before chunking) and annotate it as finite or non-finite as the case may be,

h14a. *laDZakA seba* [*khAtA*_VM *raHA*_VAUX *wA*_VAUX**]_VF**
'boy' 'apple' 'eating' 'PROG' 'was'
h15a. *seba* [*khAtA*_VM *huA*_VAUX**]_VNF** *laDZakA jA rahA thA*
'apple' 'eating' 'PROG' 'boy' ' go' 'PROG' 'was'

This approach is more faithful to the available linguistic information. However, it requires introducing another layer. So, this was not considered useful.

2b) Mark the finiteness at the chunk level,

In this approach, the lexical items are marked as in (2). No new layer is introduced. Instead, the decision is postponed to the chunk level. Since the finiteness is in the group, it is marked at the chunk level. This offers the best solution as it facilitates marking the linguistic information as it

is without having to introduce a new layer.

h14a. *laDZakA seba* ((*khAtA*_VM *raHA*_VAUX *wA*_VAUX))**_VGF**

'boy' 'apple' 'eating' 'PROG' 'was'

h15a. *seba* ((*khAtA*_VM *huA*_VAUX))**_VGNF** *laDZakA jA rahA thA*

'apple' 'eating' 'PROG' 'boy' ' go' 'PROG' 'was'

In this case also the decision is made by looking at the entire group. (2b) was most preferred as it facilitates marking the linguistic information correctly, at the same time no new layer needs to be introduced. Therefore, the current tagging scheme has adopted this approach. Thus, the main verbs in a given verb group will be marked as **VM,** irrespective of whether the total verb group is finite of non finite. Given underneath are some examples of other verb group types :

1. **Non finite verb groups -** Non-finite verb groups can have two functions :

a) Adverbial participial, for example : *khAte-khAte* in the following Hindi sentence,

h16. *mEMne* ***khAte – khAte*** *ghode ko dekhA*

'I pas' 'while eating' 'horse' 'acc' 'saw'

"I saw a horse while eating".

The main verb in (h16) would be annotated as follows :

h16a. *mEMne* ***khAte – khAte*****_VM** *ghode ko dekhA*

b) Adjectival participial, for example : '*khAte Hue*' in the following Hindi sentence ,

h17. *mEMne ghAsa* ***khAte_VM hue*** *ghoDe ko dekhA* *

'I erg' 'grass' 'eating' 'PROG' 'horse' 'acc' 'saw'

I saw the horse eating grass.

(* (h17) is ambiguous in Hindi. The other sense that it can have is, *I saw the horse while (I was) eating grass*. In such cases, the annotator would disambiguate the sentence depending on the context and mark accordingly.)

2) Gerunds

Functionally, gerunds are nominals. However, even though they function like nouns, they are capable of taking their own arguments,eg. *pInA* in the following Hindi sentence can occur on its own or take an argument (given in parenthesis):

h18. (*sharAba)* ***pInA*_VM** *sehata ke liye hAnikAraka hE.*
'liquor' 'drinking' 'heath' 'for' 'harmful' 'is'
"Drinking (liquor) is bad for health"

h19. *mujhe* ***khAnA*_VM** *acchA lagatA hai*
'to me' 'eating' 'good' 'appeals'
"I like eating"

h20. ***sunane*** ***meM*** *saba kuccha acchA lagatA hE*
'listening' 'in' 'all' 'things' 'good' 'appeal' 'is'

As mentioned above, noun '*sharAba*' in (h18) is an object of the verb '*pInA*' and has no relation to the main verb (*hE*). In order to be able to show the exact verb-argument structure in the sentence, it is essential that the crucial information of a noun derived from a verb is preserved. Therefore, even gerunds have to be marked as verbs. It is proposed that in keeping with the approach adopted for non-finite verbs, mark gerunds also as **VM** at the lexical level. For capturing the information that they are gerunds, such verbs will be marked as **VGNN** (see the section on Chunk tags for details) at the chunk level to capture their gerundial nature. The verbs having

'vAlA' vibhakti will also be marked as VM. For example, '*khonevAlA*' (one who looses).

2.2.5.5. VAUX Verb Auxiliary

All auxiliary verbs will be marked as VAUX. This tag has been adopted as such from the Penn tags. (For examples, see h14 – h16 above).

2.2.5.6. JJ Adjective

This tag is also taken from Penn tags. Penn tag set also makes a distinction between comparative and superlative adjectives. This has not been considered here. Therefore, in the current scheme for Indian languages, the tag JJ includes the 'tara' (comparative) and the 'tama' (superlative) forms of adjectives as well.
For example, Hindi *adhikatara* (more times), *sarvottama* (best), etc. will also be marked as JJ.

2.2.5.7. RB Adverb

For the adverbs also, the tag RB has been borrowed from Penn tags. Similar to the adjectives, Penn tags make a distinction between comparative and superlative adverbs as well. This distinction is not made in this tagger. This is in accordance with our philosophy of coarseness in linguistic analysis.
Another important decision for the use of RB for adverbs in the current scheme is that :-

(a) The tag RB will be used ONLY for 'manner adverbs' . Example,

h21. *vaha jaldI jaldI khA rahA thA*
'he' 'hurriedly' 'eat' 'PROG' 'was'

(b) The tag RB will NOT be used for the time and manner expressions unlike English where time and place expressions are also marked as RB. In our scheme, the time and manner expressions such as '*yahAz – vahAz, aba – waba* ' etc will be marked as PRP.

22.5.8. PSP Postposition

All Indian languages have the phenomenon of postpositions. Postpositions express certain grammatical functions such as case etc. The postposition will be marked as PSP in the current tagging scheme. For example,

h22. m*ohana kheta* ***meM*** *khAda dAla rahA thA*
'Mohan' 'field' 'in' 'fertilizer' 'put ' 'PROG' 'was'

meM in the above example is a postposition and will be tagged as PSP.
A postposition will be annotated as PSP ONLY if it is written separately. In case it is conjoined with the preceding word it will not be marked separately. For example, in Hindi pronouns the postpositions are conjoined with the pronoun,

h23. *mE****ne*** *usa****ko*** *bAzAra* ***meM*** *dekhA*
'I' 'him' 'market' 'in' 'saw'

(h23) above has three instances of 'postposition' (in bold) usage. The postpositions '*ne*' and '*ko*' are conjoined with the pronouns *mEM* and *usa* respectively. The third postposition '*meM*' is written separately. In the first two instances, the postposition will not be annotated. Such words will be annotated with the category of the head word. Therefore, the three instances mentioned above will be annotated as shown in (h23a) below :

h23a. *mE****ne*****_PRP** *usa****ko*****_PRP** *bAzAra*_**NN** ***meM*****_PSP** *dekhA*

5.9. RP Particle

Expressions such as *bhI, to, jI, sA, hI, nA*, etc in Hindi would be marked as RP.
The *nA* in the above list is different from the negative *nA.* Hindi and some other Indian languages have an ambiguous 'nA' which is used both for negation (NEG) and for reaffirmation (RP).

Similarly, the particle *wo* is different from CC *wo*. For example in Bangla and Hindi:

Bangla : (b1) *tumi* ***nA*_RP** *khub dushtu*
'you' 'particle' 'very' 'naughty'
"You are very naughty" (comment)

Hindi : (h24) *tuma* ***nA*_RP**, *bahuta dushta ho*
'you' 'particle very naughty
"You are very naughty" (comment)

Bangla : (b2) *cheleta dushtu* ***nA*_NEG**
'the boy' 'naughty' 'not'
"The boy is not naughty"

Hindi : (h25) *mEM* ***nA*_NEG** *jA sakUMgA*
'I' 'not' 'go' 'will able'
"I will not be able to go"

Bangla : (b3) *binu yYoxi khAya* ***to*_CC** *Ami khAba*
'Binu' 'if' 'eats' 'then' 'I' 'will eat'
"If Binu eats then I will eat (too)"

Hindi : (h26) *yadi binu khAyegA* **wo_CC** *mEM khAUMgI*
'if' 'Binu' 'eats' 'then' 'I' 'will eat'
"Only if Binu eats, I will eat (too)"

Bangla : (b4) *Ami* ***to*_RP** *jAni nA*
'I' 'particile' 'know' 'not'
"I don't know"

Hindi : (h27) *mujhako* ***to*_RP** *nahIM patA*
'I' 'particile' 'not' 'know'
"I don't know"

2.2.5.10. CC Conjuncts (co-ordinating and subordinating)

The tag CC will be used for both, co-ordinating and subordinating conjuncts. The Penn tag set has used IN tag for prepositions and subordinating conjuncts. Their rationale behind this is that subordinating conjuncts and prepositions can be distinguished because subordinating conjuncts are followed by a clause and prepositions by a noun phrase.

But in the current tagger all connectives, other than prepositions, will be marked as CC.

h28. mohana bAzAra jA rahA hE **Ora_CC** ravi skUla jA rahA hE
'Mohan' 'market' 'go' 'PROG' 'is' 'and' 'Ravi' 'school' 'go' 'PROG' 'is'
"Mohan is going to the market and Ravi is going to the school"

h29. mohana ne mujhe batAyA **ki_CC** Aja bAzAra banda hE
'Mohan' 'erg' 'to me' 'told' 'that' 'today' 'market' 'close' 'is'
"Mohan told me that the market is closed today."

2.2.5.11. QW Question Words

The Penn tag set makes a distinction between various uses of 'wh-' words and marks them accordingly (WDT, WRB, WP, WQ etc). The 'wh-' words in English can act as questions, as relative pronouns and as determiners. However, for Indian languages we need not keep this distinction. Therefore, we tag the question words as QW.

h30. ***kOna*** *AyA hE* ?
'who' 'come' 'has'
"Who has come ?

h31. *tuma kala* ***kyA*** *kara rahe ho* ?
'you' 'tomorrow' what' 'doing' 'are'
What are you doing tomorrow ?

h32. *tuma kala* ***kahAz*** *jA rahe ho* ?

'you' 'tomorrow' 'where' 'going' 'are'

"Where are you going tomorrow ?

h33. ***kyA*** *tuma kala Aoge* ?

'?' 'you' 'tomorrow' 'will come'

"Will you come tomorrow ?

2.2.5.12.1 QF Quantifiers

All quantifiers like Hindi *kama* (less), *jyAdA* (more), *bahuwa* (lots), etc. will be marked as QF.

h34. *vahAz* ***bahuta*_QF** *loga Aye the*

'there' many' 'people' 'came' 'was'

*"*Many people came there".

In case these words are used in constructions like '***baHutoM ne*** *jAne se inkAra kiyA*' ('many' 'by' 'to go' 'refused'; Many refues to go) where it is functioning like a noun, it will be marked as NN (noun). Quantifiers of number will be marked as below.

2.2.5.12.2 QC Cardinals

Any word denoting a cardinal number will be tagged as QC. Penn tag set has a tag CD for cardinal numbers and they have not talked of ordinals. For example,

h35. *vahAz* ***tIna*_QC** *loga bEThe the*

'there' 'three' 'people' 'sitting' 'were'

"Three people were sitting there"

2.2.5.12.3 QO Ordinals

Expressions denoting ordinals will be marked as QO.

h36. *mEMne kitAba* ***tIsare*_QO** *laDake ko dI thI*

'I' 'book' 'third' 'boy' 'to' 'give' 'was'
I gave the book to the third boy"

2.2.5.12.4 CL Classifiers

The tag CL has been included to mark classifiers. Many Indian languages have a rich classifier system. "A **classifier**, in linguistics, is a word or morpheme used in some languages to classify a noun according to its meaning" (http://en.wikipedia.org/wiki/Classifier_%28linguistics%29).

For example,

Telugu : (t2) *padi* ***mandi*** *pillalu*
'ten' 'persons' 'children'

Tamil : (tm1) *pattu* ***pEr*** *mANavarakaLa*
'ten' 'person' 'students'

The words 'mandi' (Telugu) and 'per' (Tamil) are classifiers which occur with numerals with human nouns. Such expressions when occurring separately (not suffixed with the noun) will be marked as CL. Therefore :

Telugu : (t2) *padi* ***mandi_*CL** *pillalu*
'ten' 'persons' 'children'

Tamil : (tm1) *pattu* ***pEr_*CL** *mANavarakaLa*
'ten' 'person' 'students'

2.2.5.13. INTF Intensifier

This tag is not present in Penn tag set. Words like *'bahuta*', *'kama'*, etc. when intensifying

adjectives or adverbs will be annotated as INTF. Example,

h37. *hEdarAbAda meM aMgUra* ***bahuta*_INTF** *acche milate hEM*
'HyderabAd' 'in' 'grapes' 'very' 'good' 'available' 'are'
"Very good grapes are available in Hyderabad".

2.2.5.14. INJ Interjection

The interjections will be marked as INJ. Apart from the interjections, the affirmatives such as Hindi 'HAz'('yes') will also be tagged as INJ. Since, this is the only example of such a word, it has been clubbed under Interjections.

h38. ***arre*_INJ**, *tuma A gaye !*
'oh' 'you' 'come' 'have'
"Oh! you have come"

h39. ***hAz*_INJ**, *mEM A gayA*
'yes', 'I' 'come' 'have'
"Yes, I have come".

5.15. NEG Negative

Negatives like Hindi 'nahIM' (not), 'nA' (no, not), etc. will be marked as NEG. For example,

h40. *Vaha Aja* ***nahIM*_NEG** *A pAyegA*
'he' 'today' 'not' 'come' 'will be able'

Also, see examples (b2) and (h25) given above.
Indian languages have reiteration of NEG in certain constructions. For example,

b5. T*umi chobitA dekhbe* ?
'you' 'picture-def' 'will see' ?
"Will you see the picture ?"

b6. *nA*_NEG, *xekhabo nA*_NEG
*'n*o' 'will see (I)' 'not'
"No, I will not see (it)"

The first occurrence of *'nA'* in such constructions will also be marked as NEG.

2.2.5.16. UT Quotative

A quotative introduces a quote. Typically, it is a verb. Many Indian languages use quotatives. Given below is an example from Bengali,

b7. *She Ashbe* ***bole*** *bolechilo*
'he' 'will come' 'quotative' 'told'
"He told that he will come".

2.2.5.17. SYM Special Symbol

All those words which cannot be classified in any of the other tags will be tagged as SYM. This tag is similar to the Penn 'SYM'. Also special symbols like $, %, etc are treated as SYM. Since the frequency of occurrence of such symbols is very less in Indian languages, no separate tag is used for such symbols.

5.18. *C Compounds (Make it XC – where X is a variable of the type of the compound of which the current word is a member of)

The issue of including a tag for marking compounds was discussed extensively. Results of

algorithms using IIIT-H tag set which included NNC (part of compound nouns) and NNPC (part of proper nouns) showed that these two tags contributed substantially to the low accuracy of the tagger. Since most elements which occur as NNC or NNPC can also occur as NN and NNP, it affected the learning by the machine. So, the question was, why to include tags which contributed more to the errors ? The other aspect, however, was that while human annotators are annotating the data, they know from the context when a certain element is NNC or NN, NNPC or NNP and if marked, this information can be useful for certain applications. The argument is same as the one in favor of including a tag for proper nouns.

Another point which was discussed was that any word class can have compound forms in Indian languages (including adjectives and adverbs).

Therefore, if we decide to have a tag for showing compounds of each type, the number of tags will go very high. The final decision on this was to include a *C tag which will be ccur n as **catC** tag of the type of compound that the element is a part of. For example, if a certain word is part of a compound noun, it will be marked as NNC, if it is part of a compound adjective, it will be marked as JJC and so on and so forth.

Some examples are given below :

Hindi compound noun *keMdra sarakAra* (Central government) will be tagged as *keMdra*_**NNC** *sarakAra*_**NN**.

In this example, '*keMdra*' and '*sarakAra*' are both nouns which are forming a compound noun. All words except the last one, of a compound words will be marked as NNC. Thus any NNC will be always followed by another NNC or an NN. This strategy helps identify these words as one unit although they are not conjoined by a hyphen. Similarly, a compound proper noun will be marked as NNPC excluding the last one. Eg. *aTala*_NNPC *bihArI*_NNPC *vAjapeyI*_NNP

The first two words, in the above example, will be tagged as NNPC and the last one will be tagged as NNP. Similar to the NNC tag for common nouns, NNPC tag helps in marking parts of a proper noun.

H41. *rAma, mohana aur shyAma ghara gaye*.

'Ram', 'Mohan' 'and' 'Shyam' 'home' 'went'

"Ram, Mohan and Shama went home".

H42. *bagIce meM* ***ranga*_JJC** ***biraMge*_JJ** *phUla khile the*
'garden' 'in' 'colourful' 'flowers' 'flowered' 'were'
"The garden had colorful flowers"

Titles such as **Dr., Col., Lt**. etc. which may occur before a proper noun will be tagged as **NNC**. All such titles will always be followed by a Proper Noun. In order to indicate that these are parts of proper nouns but are nonetheless nouns themselves, they will be tagged as NNC, eg. **Col._NNC** Ranjit_**NNPC** Deshmukh**_NNP**

2.2.5.19 RDP Reduplication

In this phenomenon of Indian languages, the same word is written twice for various purposes such as indicating emphasis, deriving a category from another category etc. eg. *choTe choTe* ('small' 'small'; very small), *lAla lAla* ('red' 'red'; red), *jaldI jaldI* ('quickyl' 'quickly' ; very quickly)
There are two ways in which such word sequences may be written. They can be written – (a) separated by a space or (b) separated by a hyphen.

The question to be resolved is that in case, they are written as two words (separated by space)– how should they be tagged? Earlier decision was to use the same tag for both the words. However, in this approach, the morphological character of reduplication is missed out. That is, the reduplicated item will then be treated exactly like two independent words of the same category. For example,

h43. *Vaha* ***mahaMgI*_JJ** ***mahaMgI*_JJ** *cIjZeM kharIda lAyA*
'he' 'expensive' 'expensive' 'things' 'buy' 'bring'
"He bought **all expensive** things".
H44. *Una* ***catura_JJ buddhimAna_JJ*** *baccoM ne samasyA sulajhA lI*

‘those’ ‘smart’ ‘intelligent’ ‘children’ ‘erg’ ‘problem’ ‘solved’

“Those **smart** and **intelligent** children solved the problem.

Both (h43) and (h44) have a sequence of adjectives - *mahaMgI*_JJ *mahaMgI*_JJ and *catura_JJ buddhimAna_JJ* respectively. In the first case, the sequence of two adjectives is a case of reduplication (same adjective is repeated twice to indicate the intensity of ‘expensive’) whereas in the second case the two adjectives refer to two different properties attributed to the following noun. Since reduplication is a highly productive process in Indian languages, it is proposed to include a new tag **RDP** for annotating reduplicatives. The first word in a reduplicative construction will be tagged by its respective lexical category and the second word will be tagged as RDP to indicate that it is a case of reduplication distinguishing it from a normal sequence such as in (h44) above. Some more examples are given underneath to make it more explicit,

h45. *Vaha **dhIre**_**RB** **dhIre**_**RDP** cala rahA thA.*

‘he’ ‘slowly’ ‘slowly’ ‘walk’ ‘PROG’ ‘was’

“He was walking (very) slowly”.

H46. *Usake bAla **choTe**_**JJ** **choTe**_**RDP** the.*

‘his’ ‘hair’ ‘short’ ‘short’ ‘were’

“He had (very) short hair”

h47. Yaha bAta ***galI***_**NN** ***galI***_**RDP** *meM phEla gayI.*

‘this’ ‘talk’ ‘lane’ ‘lane’ ‘in’ ‘spread’ ‘went’

“The word was spread in every lane”.

2.2.5.20. ECHEcho words

Indian languages have a highly productive usage of echo words such as Hindi ‘*cAya-vAya*’ (‘tea’ ‘echo’), where *‘cAya*’ is a regular lexical item of Hindi vocabulary and *‘vAya*’ is an echo word indicating the sense “etc” . These words, on their own, are ‘nonsense’ words and do not find a place in any dictionary.

Thus, the gloss for ‘*cAya-vAya*’ would be ‘*tea etc*’. It is proposed to add the tag **ECH** for such

words.

2.2.5.21. UNKUnknown

A special tag to indicate unknown words is also included in the tag set. The annotators can use this tag to mark the words whose category they are not aware of. This tag has to be used very cautiously and sparsely, i.e., only if it is absolutely necessary.

2.2.6.Some Special Cases

This section gives the details of ccur n aspects of Indian languages which need to be dealt with separately in the tagger. These are issues that cannot be handled by just changing or adding tags.

2.2.6.1 'vAlA' type constructions

'*vAlA*' is a kind of suffix used in Hindi and some other Indian languages. It conjoins with nouns (Case I, below) or verbs (Case II) to form adjectives or even nouns. It is also used as an aspectual TAM in a verbal construction (Case III).

H48. *lAla* ***kamIjZa vAlA*** *AdamI merA bhAI hE* .
'red' 'shirt' 'in' 'man' 'my' 'brother' 'is'
"The man in red shirt is my brother".

H49. *Mehanawa* ***karane vAle*** *vyakti ko inAma milegA* .
'hard work' 'doing' 'adj' 'person' 'to' 'prize' 'will get'
The person who works hard will get a prize.

These cases are elaborated below.

Case I: The suffix 'vAlA' can ccur with a noun. For example, ***lAThI vAlA (***

'stick' 'with' –The one with a stick).

H50. *lAThI vAle AdamI ko bulAo*
'with stick' 'man' 'acc' 'call'
"Call the man with the stick".

This suffix *'vAlA*' in Hindi (a) may be written separately or (b) may be attached to the preceding noun.

(c) In case it is written separately as in '*lAThI vAlA*' above, the word '*lAThI*' will be tagged as NN and the word '*vAlA*' will be tagged as PSP.

The whole expression '*lAThI vAlA*' is an adjective, in which '*lAThI*' is a noun and '*vAlA*' is a suffix which derives an adjective from a noun. Since '*lAThI*' and '*vAlA*' written separately in the above example, they have to be tagged individually. '*vAlA*' in such cases will be treated like a postposition and will be tagges as PSP.

(b) The second possibility is of '*lAThi*' and '*vAlA*' written together as '*lAThIvAlA*'. In such cases it will be treated as one word and will be marked as JJ since 'l*AThIvAlA*' is an adjective.

Case II: '*vAlA*' can also occur after a verb. Example, ***karane vAlA*** ('doing' 'one' – The one who does something)

h51. *Mehanata* ***karanevAle ko*** *phala milatA hE*
'hard' 'working one' acc 'fruit' 'get' 'PRES'
The one who works hard gets the fruits".

As mentioned earlier, the suffix '*vAlA*' also joins a verb in its nominal form and makes it an adjective. In this case also, the two words may be written separately (*karane vAle*) or together

(*karanevAlA*). In the former case, the two words will be marked as VM and PSP respectively (*karane*_VM *vAle*_PSP). In the latter case, being a single word (*karanevAlA*) it will be tagged as VM (*karanevAle*_VM). It is crucial to retain the 'verb' information in these case, so that at a later stage if we want to annotate its argument structure we should be able to do so (discussed earlier in the document).

Case III: 'vAlA' can also occur as part of TAM. For example,

h52. *mEM wo* ***jAne*** ***vAlA*** *hI* *thA.*
'I' 'particle' 'to go' 'about' 'part' 'was'
"I was about to go"

Although the word '*jAne*' has a '*vAlA*' suffix in (h52) above, the entire expression is not an adjective but is a verb having the aspectual information of 'shortly'. In this case, the sequence '*jAne vAlA*' will be marked as *jAne*_VM
*vAlA*_VAUX. The alternative writing convention of writing the sequence as one word (*jAnevAlA*) is possible in this case also. Like the earlier cases, the word will be marked for the category of the content morpheme – which is verb in this case. Thus *jAnevAle* will be tagged as *jAnevAle*_VM.

Here again we stand by our policy that the tag will be decided on the basis of the part of speech and not on the basis of the category of the word in the given sentence(syntactic function). This avoids confusion at the level of manual tagging and aids machine learning as well. So the tag (VM) remain same although the function of the words is different in two different places, it is adjective in Cases I and II and verbal in Case III.

2.2.6.2 Honorifics in Indian languages

Hindi (and some other Indian languages) has particles such as '*jI*' or '*sAHaba*' etc. after proper nouns or personal pronouns. These particles are added to denote respect to the referred person. Such honorific words will be treated like particles and will be tagged **RP** like other particles.

H53. ***mantrI*_NN *jI*_RP** *sabhA meM dera se pahuMce* .

'minister' 'hon' 'meeting' 'in' 'late' 'part' 'reached'

"The minister reached late for the meeting".

2.2.6.3 Foreign words

Presence of loan words is a fairly common phenomenon in languages. Most Indian languages have a number of loan word from English. One may also come across words from other Indian languages or Sanskrit in a given text. Such foreign words will be tagged as per the syntactic function of the word in the given context. In special cases, such as when the annotator is not sure of the category of a word, it will be tagged as **UNK.**

2.2.7. A Special Note

There may be situations, when an annotator does not feel very confident about the tag for a particular word. The annotator may then assign it different tags in different places. Inconsistency in the manual tagging can affect the learning considerably. Since this is a task which involves a number of human annotators, the methods have to be evolved to check and cross validate the human annotation. Another practical problem in annotation is that in the initial stages of annotation, the annotators need time to get familiar with the tagging scheme and the concept behind each tag. Thus they take some time before coming to a stable stage of decision making for various instances, particularly various ambiguous cases. Especially, in the initial stages, the annotators may often come across cases where their confidence level may not be very high. They may feel the need of some clarifications for these cases. Since the task of annotation has to go on and immediate clarification may not be possible, the annotators may be forced to take decisions and mark a case as they consider appropriate at that point of time. Over a period of time, with better understanding of the tags and tagging scheme, they may reach a stable stage. However, by then they may already have tagged a given case differently in different places thus introducing inconsistency in the annotated corpus. At a later stage, it will be difficult to go back to all the

cases that have been annotated by then and correct them. So the chances are that the annotators may proceed with the revised decision and leave the earlier annotation as such. This will introduce inconsistencies in the annotated corpus.

To control such a situation, it is decided to provide a way by which the annotators can initially mark the uncertainty of their decision so that they can easily extract these cases easily and take them up for discussions and clarifications.

This 'uncertainty' will be annotated as follows:

c) The annotators first mark such a case with a tag that they consider appropriate at the time of annotation.

b) Along with the chosen tag, they also put a question mark (?) against that tag. The question mark will indicate that this case is not yet resolved and will be finalized after clarification or discussion.

c) All the cases with a question mark can be later taken out and placed for discussion. An annotator will be responsible for bringing such cases for discussion and once the cases are resolved, the annotator will go back and correct the tag. In case the tag assigned by the annotator initially itself is correct, the annotator will remove the question mark against it.

This is a purely temporary measure and the data finally submitted by an annotator should not have any words having a question mark.

3. CHUNKING

3.1. Chunk Tags Chosen for the Current Scheme

This section deals with the chunk tags. Not many of the issues discussed above hold for defining the chunk tags. Various points which have been deliberated upon in relation to chunking scheme are :

1. 1. Definition of a chunk
2. 2. Chunk Types
3. 3. Some Special Cases
4. 4. Annotation method/procedure

3.2. Definition of a chunk

Following issues related to the definition of a chunk were discussed :

What constitutes a 'chunk' ?

A typical chunk consists of a single content word surrounded by a constellation of function words (Abney,1991). Chunks are normally taken to be a 'correlated group of words'.

The next issue, however, is - How to define the boundaries of these 'correlated word groups' for our purpose?

For example, which case in the following pairs should be grouped as a chunk ?

((*xillI meM*)) OR ((*xillI*)) *meM*
'Delhi' 'in' *'Delhi' 'in'*
((*rAjA kA betA*)) OR ((*rAjA kA*)) ((*betA*))
'king' 'of' 'son' 'king 'of' 'son'
((*rAjA ke bete kI paxnI*)) OR ((*rAjA ke*)) ((*bete k*I)) ((*paxnI*))
'king' 'of' 'son' 'of' 'wife' 'king' 'of' 'son' 'of' 'wife'

Following definition of a 'chunk' was evolved through discussion :

“A minimal (non recursive) phrase(partial structure) consisting of correlated, inseparable words/entities, such that the intra-chunk dependencies are not distorted”. Each chunk type discussed and the decided upon is described below **.**

3.3. Chunk Types

Based on the above definition of chunk, issues related to various chunk types were discussed. A chunk would contain a 'head' and its modifiers.

3.3.1 NP Noun Chunk

Noun Chunks will be given the tag NP and include non-recursive noun phrases and postpositional phrases. The head of a noun chunk would be a noun. Specifiers will form the left side boundary for a noun chunk and the vibhakti or head noun will mark the right hand boundary for it. Descriptive adjective/s modifying the noun will be part of the noun chunk. The particle which anchors to the head noun in a noun chunk will also be grouped within the chunk. If it occurs after the noun or vibhakti, it will make the right boundary of the chunk.

Some example noun chunks are :

((*bacce*_NN))_NP, ((*kucha*_QF *bacce*_NN))_NP,

'children' 'some' 'children'

((*kucha*_QF *acche*_JJ *bacce*_NN))_NP, ((*Dibbe*_NN *meM*_PSP))_NP,

'some' 'good' 'children' 'box' 'in'

((eka_QC *kAlA*__JJ *ghoDZA*_NN))_NP ,

'one' 'black' 'horse'

((*yaha*_DEM *nayI*_JJ *kitAba*_NN))_NP,

'this' 'new' 'book'

((**isa**_DEM *nayI*_JJ *kitAba*_NN *meM*_PREP))_NP,

'this' 'new' 'book' 'in'

((*isa*_DEM *nayI*_JJ *kitAba*_NN *meM*_PSP *bhI*_RP))_NP

'this' 'new' 'book' 'in' 'also'

The issue of genitive marker and its grouping with the nouns that it relates to was discussed in detail. For example, the noun phrase '*rAma kA beTA*' contains two nouns '*rAma*' and '*beTA*'. The two nouns are related to each other by the vibhakti '*kA*'. The issue is whether to chunk the two nouns separately or together? Linguistically, '*beTA*' is the head of the phrase "*rAma kA beTA*".

'*rAma*' is related to '*beTA*' by a genitive relation which is expressed through the vibhakti '*kA*'. Going by our definition of a 'chunk' we should break '*rAma kA beTA*' into two chunks (((*rAma kA*))_NP, ((*beTA*))_NP) by breaking '*rAma kA*' at 'kA' vibhakti . Moreover, if we chunk 'rAma kA beTA' as one chunk, linguistically, we will end up with a recursive noun phrase as a single chunk ((((*rAma kA)) beTA*)) which also is against our definition of a chunk.

Therefore, it was decided that the genetive markers will be chunked along with the preceding noun. Thus, the noun group 'rAma kA beTA' would be chunked into two chunks.

h54. **((*rAma kA*))NP** ((*beTA*))NP acchA hE "Ram's son is good"

h55. ((*kitAba*))NP **((*rAma kI*))NP** hE "The book belongs to Ram"

For the noun groups such as "*usakA beTA*" it was decided that they should be chunked together.

3.3.2 Verb Chunks

The verb chunks would be of several types. A verb group will include the main verb and its auxiliaries, if any. Following are some examples of verb chunks from Hindi,

((*khAyA*)), ((*khA rahA hE*)), ((*khA sakawe hEM*))

'ate' 'eat' 'PROG' 'is' 'eat' 'can' 'PRES'

The types of verb chunks and their tags are described below.

3.3.2.1 VGF Finite Verb Chunk

As mentioned in 5.4 above, a verb group sequence (V VAUX VAUX . .) contains a main verb and its auxiliaries. The group itself can be finite or non-finite. In case of it being finite, the main verb in such a sequence may not be finite. The finiteness is known by the auxiliaries. Therefore,

it is decided to mark the finiteness of the verb at the chunk level. Thus, any verb group which is finite will be tagged as **VGF.** For example,

h56. *mEMne ghara para khAnA ((khAyA*_VM*))*_**VGF**

'I erg' 'home' 'at' 'meal' 'ate'

h57. *vaha cAvala ((khA*_VM *rahA*_VAUX *hE*_VAUX*))*_**VGF**

'he' 'rice' 'eat' 'PROG' 'is'

3.3.2.2.2 VGNF Non-finite Verb Chunk

A non-finite verb chunk will be tagged as **VGNF.** For example,

h15a) *seba ((**khAtA**_**VM** **huA**_**VAUX))_**VGNF** laDZakA jA rahA thA*
'apple' 'eating' 'PROG' 'boy' ' go' 'PROG' 'was'

h16a) *mEMne ((**khAte – khAte**_**VM))**_**VGNF** ghode ko dekhA*

'I erg' 'while eating' 'horse' acc 'saw'

h17a) *mEMne ghAsa ((**khAte**_**VM** **hue**_**VAUX))_**VGNF** ghoDe ko dekhA*
'I erg' 'grass' 'eating' 'PROG' 'horse' acc 'saw'

The IIIT-H tagset had initially included three tags for the non-finite verbal forms. Unlike Penn tagset, all non finite verbs, which are used as adjectives, were marked as VJJ at the POS level. Similarly, to mark adverbial non-finite verbs, the POS tagset had VRB tag. A tag VNN was included to mark the nominalized verbs.

However, during the discussions IL standards, it was pointed out that inclusion of too many finer tags hampers machine learning. Moreover, the marking is based on syntactic information, which we should avoid at the POS level, unless it is contributing to further processing in a substantial way. On the other hand, it is important to mark finite non-finite distinction in a verbal expression as it is a crucial information and is also easy to learn. As discussed under 5.4 above, it was

decided to mark this distinction at the chunk level, rather than at the POS level. Therefore, the tag VGNF has been included to mark non-finite adverbial and adjectival verb chunk.

3.3.2.2.3 VGINF Infinitival Verb Chunk

This tag is to mark the infinitival verb form. In Hindi, both, gerunds and infinitive forms of the verb end with a *-nA* suffix. Since both behave functionally in a similar manner, the distinction is not very clear. However, languages such as Bangla etc have two different forms for the two types. Examples from Bangla are given below.

b8. *Borabela **((snAna karA))_*****VGNN** *SorIrera pokze BAlo*
'Morning' 'bath' 'do-verbal noun' 'health-gen' 'for' 'good'
'Taking bath in the early morning is good for health"

b9. *bindu Borabela ((**snAna karawe))_***VGINF** *BAlobAse*
'Bindu' 'morning' 'bath' 'take-inf' 'love-3pr'
"Bindu likes to take bath in the early morning"

In Bangla, the gerund form takes the suffix *–A / -Ano*, while the infinitive marker is *–we*. The syntactic distribution of these two forms of verbs is different. For example, the gerund form is allowed in the context of the word *darakAra* "necessary" while the infinitive form is not, as exemplified below:

b10 *Borabela ((snAna karA))_***VGNN** *darakAra*
'Morning' 'bath' 'do-verbal noun' 'necessary'
"It is necessary to take bath in the early morning"

b11. **Borabela ((snAna karawe))_***VGINF** *darakAra*

Based on the above evidence from Bangla, the tag *VGINF* has been included to mark a verb

chunk.

3.2.2.2.4 VGNN Gerunds

A verb chunk having a gerund will be annotated as **VGNN**. For example,

h18a. *sharAba ((**pInA**_**VM))_VGNN** sehata ke liye hAnikAraka hE.*

'liquor' 'drinking' 'heath' 'for' 'harmful' 'is'

"Drinking (liquor) is bad for health"

h19a. *mujhe rAta meM ((**khAnA**_**VM))_VGNN** acchA lagatA hai*

'to me' 'night' 'in' 'eating' 'good' 'appeals'

"I like eating at night"

h20a. ((***sunane***_**VM** ***meM***_**PSP))_VGNN** *saba kuccha acchA lagatA hE*

'listening' 'in' 'all' 'things' 'good' 'appeal' 'is'

3.2.2.3 JJP Adjectival Chunk

An adjectival chunk will be tagged as **JJP**. This chunk will consist of all adjectival chunks including the predicative adjectives. However, adjectives appearing before a noun will be grouped together with the noun chunk. A JJP will consist of phrases like

h58. *vaha laDaZkI hE*((*suMdara*_**JJ** *sI*_RP))_**JJP**

'she' 'girl' 'is' 'beautiful' 'kind of'

h59. *hAthI AyA* ((*moTA*_*C *tagadA*_**JJ**))_**JJP**

'elephant' 'came' 'fat' 'powerful'

h60. *vaha laDakI ((bahuta*_INTF *sundara*_**JJ**))_**JJP** *hE*

'she' 'girl' 'very' 'beautiful' 'is'

Cases such as (h61) below will not have a separate JJP chunk. In such cases, the adjectives will be grouped together with the noun they modify. Thus forming a NP chunk.

h61. ((*kAle*_**JJ** *ghane*_**JJ** *laMbe*_**JJ** *bAla*_**NN**))_**NP**

'black' 'thick' 'long' 'hair'

3.2.2.3.1 Some special cases

Following examples from Hindi present a

h62. *xillI meM* ***rahanevAlA*** *merA BAI kala A rahA hE .*

'Delhi' 'in' 'staying' 'my' 'brother' 'tomorrwo' 'come' 'PROG' 'is'

"My brother who stays in Delhi is coming tomorrow".

h63. *usane Tebala para* ***rakhA huA*** *seba khAyA.*

'(s)he erg' 'table' 'on' 'kept' 'apple' 'ate'

"He ate the apple kept on the table".

In (h62) above '*rahanevAlA*' is an adjectival participle. But we do NOT mark it as JJP. Instead, it will be marked as a **VGNF**. The decision to tag it as a VGNF is based on the fact that such adjectival participles are derived from a verb can have their arguments. This information is useful for processing at the syntactic level. Thus, '*rahanevAlA*' in (h62) will be annotated as follows:

h62a. *xillI meM* ((***rahanevAlA*_VM*)*_VGNF** *merA BAI kala A rahA hE .*

Similarly, in (h63) above, the chunk '*rakhA huA*' is an adjective but will also be marked as a VGNF since this also derived from a verb and chunks like '*Tebala pra*' etc are its arguments. So the chunk name will be **VGNF** and the POS tag will be **VM** which might be followed by an auxiliary verb tagged as **VAUX**. (h63a) shows how '*rakhA huA*' will be annotated :

h63a. *usane Tebala para ((**rakhA**_VM **huA**_VAUX*))*_VGNF**seba khAyA.*

3.2.2.4 RBP Adverb Chunk

This chunk name is again in accordance with the tags used for POS tagging. This chunk will include all pure adverbial phrases.

h64. *vaha* ((*dhIre-dhIre*_**RB**))_**RBP** *cala rahA thA.*

'he' 'slwoly' 'walk' 'PROG' 'was'

"He was walking slowly"

Now consider the following examples:

h65. *vaha* ***dagamagAte hue*** *cala rahA thA* .

'he' ' 'walk' 'PROG' 'was'

"He was walking

h66. *vaha khAnA* ***khAkara*** *ghara gayA* .

'he' 'meal' 'after eating' 'home' 'went'

"He went home after eating his meal"

In the above examples, '*dagamagAte hue*' and 'k*hAkara*' are non finite forms of verbs used as adverbs. Similar to adjectival participles these will also be chunked as **VGNF** and not as **RBP**. The reason for this is that we need to preserve the information that these are underlying verbs. This will be a crucial information at the level of dependency marking where the arguments of these verbs will also be marked.

((isa_PRP nayI_JJ kitAba_NN meM_PSP bhI_RP))_NP

'this' 'new' 'book' 'in' 'also'

3.2.2.5 NEGP Negatives

1. In case a negative particle occurs around a verb, it is to be grouped within verb group. For example,

h67. *mEM kala dillI ((**nahIM**_**NEG** **jA**_**VM** **rahI**_**VAUX**))_**VGF**

"I" "tomorrow" "Delhi" "not" "go" "Cont"

h68. ((***binA***_**NEG** ***bole***_**VM**))_**VGNF** *kAma ((**nahIM**_**NEG** **calatA**_**VM**))*_**VGF**

"without" "saying" "work" "not" "happen"

However ,

h69. **binA** kucha **bole** kAma **nahIM calatA**

"without" "something" "saying" "work" "not" "happen"

In the above sentence, the noun "*kucha*" is coming between the negative "*binA*' and verb "*bole*". Here, it is not possible to group the negative and the verb as one chunk. At the same time, "*binA*" cannot be grouped within an NP chunk, as functionally, it is negating the verb and not the noun. To handle such cases an additional **NEGP** chunk is introduced. If a negative occurs away from the verb chunk, the negative will be chunked by itself and chunk will be tagged as NEGP. Thus,

h69a. **((*binA*))_NEGP ((***kucha*))_NP **((*bole*))_VG ((***kAma*))_NP ((*nahIM calatA***))_VG**

3.2.6 CCP Conjuncts

Conjuncts are functional units information about which is required to build the larger structures. Take the following examples of cunjunct usages :

h70. (*rAma kitAba paDha rahA thA)* ***Ora*** *(mohana Tennisa khela rahA thA).*

"Ram was reading a book **and** Mohan was playing tennis"

h71. (*rAma ne batAyA)* ***ki*** *(usakI kitAba acchI hE).*

"Ram said **that** his book is good"

h72. (*rAma)* **Ora** *(mohana) Tennisa khela rahe the*.

"Ram **and** Mohan were playing tennis".

h73. (*merA bhAI rAma) **Ora** (usakA dosta mohana) Tennisa khela rahe the*.

"My brother Ram **and** his friend Mohan were playing tennis".

h74 . *rAma (saphZeda kapade) **Ora** (nIle jute) pahane thA*.

"Ram was wearing white clothes and blue shoes".

h75. *rAma eka (halkI) **Ora** (nIlI) bOla lAyA*.

"ram brought a light **and** blue ball".

The sentences above have various types of conjoined structures. To represent these conjoined structures, it is decided to form separate chunks for conjuncts and the elements a conjunct conjoins. Thus (h70) and (h71) above will be chunked as (h70a) and (h71a) given below,

h70a. ((*rAma*))_NP ((*kitAba*))_NP ((*paDha rahA thA*))_VG **((*Ora*))CCP** ((*mohana*))_NP ((*Tennisa*))_NP ((*khela rahA thA*))_VG.

h71a. ((*rAma ne*))_NP ((*batAyA*))_NP **((*ki*))_CCP** ((*usakI*))_NP ((*kitAba*))_NP ((*acchI*))_JJP ((*hE*))_VG.

Expression '*rAma Ora mohana*' in example (h72) is a complex NP. Though complex, the expression can be annotated as a single NP chunk as functionally it is the subject of the verb 'play'. However, example (h73) presents a case where it would be better to form three independent chunks for the complex subject NP. Though the conjunct '***Ora***' is conjoining '*rAma*' and '*mohana*', both '*rAma*' and '*mohana*' have their respective modifiers. To make it explicit, it is better to treat them as two independent NP chunks conjoined by a CCP.

h73a. ((*merA bhAI rAma*)) **((*aura*))_CCP** ((*usakA dosta mohana*))_NP ((*Tennisa*))_NP ((*khela rahe the*))_VG.

Following this, the subject NP of (h72) would also be annotated similarly. Therefore,

h72a. ((*rAma*))_NP **((*aura*))_CCP** ((*mohana*))_NP ((*Tennisa)*)_NP ((*khela rahe the*))_VG.

The annotation for cases such as (h74) and (h75) would be as follows :

h74a. ((*rAma*))_NP ((*safeda kapade*))_NP **((*aura*))_CCP** ((*nIle jute*))_NP ((*pahane thA*))_VG.

h75a. ((*rAma*))_NP ((*eka*))_JJP ((*halkI*))_JJP **((*aura*))_CCP** ((*nIlI*))_JJP ((*bOla*))_NP ((*lAyA*))_VG

Thus the decision for conjuncts is - the conjoined entities will be broken into separate chunks. eg. ((*rAma*))_NP ***((Ora))*_CCP** ((*SyAma*))_NP

3.2.8.2.7 FRAGP Chunk Fragments

Some times certain fragments of chunks are separated from the chunks to which they belong. For example :

h76. ***rAma*** (*jo merA baDZA bhAI hE*) ***ne*** *kahA* ...

'Ram' 'who' 'my' 'elder' 'brother' 'is' 'erg' 'said'

In the above example, vibhakti '*ne*', which is a case marker of the noun '*rAma*', is separated from it by an intervening clause. Syntactically, '*ne*' is a part of the noun chunk '*rAma ne*'. However, at times it can be written separately. The following was decided for such fragments :

(i) There will be a separate chunk for the vibhakti in constructions where it gets separated from the noun it would normally be grouped with. This chunk can have more than one entity within it.

h77. ((***rAma*))_NP,** *mere dillI vAle bhAI,* **((*ne*))_FRAGP** *kahA*

'Ram' 'my' 'Delhi' 'from' 'brother' 'erg' 'said'

(ii) If the entities embedded between the noun and it's vibhakti are a series of nouns the entire group will be chunked as a single noun chunk.

h78. **((***isa **'upanyAsa samrATa'** Sabda kA***))_NP**

'this' 'Novel' 'King' 'word' 'of'

3.2.2.8 BLK Miscellaneous entities

Entities such as interjections and discourse markers that cannot fall into any of the above mentioned chunks will be kept within a separate chunk.

eg. ((*oh*_INJ))_**BLK,** ((*arre*_INJ))_**BLK**

3. Some Special Cases

Apart from the above, some special cases related to certain lexical types are discussed below.

3.2.2.9.Conjunct Verbs

The issue whether to treat the noun/adjective which is part of a conjunct verb differently by marking it with a special tag (NVB/JVB) or to treat it as a noun like any other noun at the POS level was deliberated on.

The question was based on the following observations :

a) NVB/JVB , as part of conjunct verbs, are most often not recognized by the learning algorithms.

b) Having NVB at the POS level is based on syntactic considerations. Therefore, do we really need to go for it ? Instead, at the POS level we mark the noun as a noun and leave the decision of marking a conjunct verb as single unit for a later level.

c) Moreover, since the noun, which is part of a conjunct verb (Kriyamula), can occur away from its 'verbaliser', it becomes difficult to differentiate it from a 'noun' which may be an argument of the verb. This also creates problem for chunking of the verb group. The two components of the chunk have to be separately marked and have to be joined at the syntactic level.

d) If NVB is marked at the POS level, a natural consequence would be to group it with its verbaliser as a VG chunk. In fact, that is the purpose of identifying it as different from a noun. However, sometimes one comes across expressions such as '*mEMne unase eka **prashna kiyA***' (I posed a question to him). In this sentence, '***eka***' is a modifier of '*prashna*'. '*prashna karana*' is recognized as a conjunct verb in Hindi by most Hindi speakers. Following example shows the problem of grouping '*praSna karanA*' as a single VG:

POS : *mEMne*_PRP *unase*_PRP *eka*_QC *prashna*_NVB *kiyA*_VM
Chunk : ((*meMne*_PRP))_NP ((*unase*_PRP))_NP ((*eka*_QC))_JJP ((*prashna*_NVB *kiyA*_VM))_VGF

Once "*praSna karanA*" are grouped together as a chunk, it will be difficult to show the relation between '*eka*' and 'prashna' subsequently.

Thus, an alternative was proposed wherein, the noun of the conjunct verb is tagged as NN at the POS level which is accordance with the decision to tag the lexical item based on its lexical category. Thereafter, the noun is grouped with its preceding adjectival modifiers as an NP chunk. The only problem in this approach is that the information of a noun verb sequence being a conjunct verb is not captured till the chunk level and the noun of the conjunct verb is separated from its verbaliser. However, the approach has following advantages :

1) At the POS level, the word is tagged for its grammatical category and not for its syntactic function. This eases the decision making at the POS level. And marking the information, that the conjunct verbs which are composed of two words form one lexeme semantically, is postponed to a later level.

2) It allows us to show the modifier-modified relation between an adjective such as '*eka*' in the above example with its modified noun '*praSna*'.

3) Since the information of a noun verb sequence being a 'kriyamula' is crucial at the syntactic

level, it will be captured at that level by marking the relation between the 'noun' and its verbaliser by an appropriate tag. Therefore, the decision is :

The noun/adjective and verb (internal components of a conjunct verb) will be chunked separately.

eg. *prashna karanA* - **((*prashna*))NP ((*kiyA*))VG**

ucita kiyA - **((*ucita*))JJP ((*kiyA*))VG**

3.2.3.2 Particles

Regarding the particles, it was decided that the particles will be chunked with the same chunk as the anchor word they occur with. Thus,

eg. ((*rAma ne* ***bhI***))_**NP, ((***mEM* ***wo*****))_NP,**

'Ram' 'erg' 'also' 'I' 'emph'

3.2. 3.3 Quantifiers

The issue of chunking quantifiers was discussed in great details. Numbers can occur (a) as noun modifiers before a noun (***haZaroM*** *ladakoM ne – 'thousands' 'boys' 'erg'*) or (b) can occur without a noun (***hazAroM*** *ne – 'thousands' 'erg'*) with a nominal inflection. The issue of whether to treat the quantifiers of the type (b) as nouns was discussed. The issue is whether (b) is a case of an ellipsis of the noun after a number or whether it is the number itself which is the noun. If the latter has to be followed then the POS tag for quantifiers in such cases should be NN. Following decisions were taken :

(i) A 'QC' or 'QO' occuring with a noun will be part of the noun chunk.

h79. ((***hazAroM*****_QC** *logoM_NN ne*_PSP))_NP *yaha driSya dekhA*

'thousands' 'people' 'erg' 'this' 'scene' 'watched'

"Thousands of people watched this scene".

h80. ((***dUsare*_QO** *ladake*_NN *ne*_PSP))_NP *isa samasyA ko sulajhA diyA*

'second' 'boy' 'erg' 'this' 'problem' acc 'solve' 'did'

"The second boy solved this problem".

(ii) All categories occurring without a noun, with nominal inflections (overt or otherwise) will be tagged as noun.

h81. ((***hazAroM_NN*** *ne*_PSP))_NP *yaha driSya dekhA*

'thousands' 'erg' 'this' 'scene' 'watched'

"Thousands watched this scene.

h82. **((*mote*_NN *ne*_PSP))_NP ((chote_NN ko_PSP))_NP ((***mArA***))_**VGF

'fat' 'erg' 'small' 'to' 'killed'

3.2. 3.4 Punctuations

All punctuations, with an exception of sentence boundary markers and clausal conjuncts, will be included in the preceding chunk. For example

h83. ((*usane*_PRP))_NP ((***kahA*_VM –_SYM)_**VGF

'He erg' 'said'

((**"_SYM *yaha*_PRP**))_NP ((*Thika*_JJ))_JJP ((***hE*_VM "_SYM**))_VGF

'this' 'proper' 'is'

"He said, "this is not right" ".

h84. *rAma AyA* **((,_SYM))_CCP** *mohana gayA*

'Ram' 'came' , 'Mohan' 'went'

"Ram came and Mohan left".

Punctuations such as (a) hyphens and (b) quote marks will be taken care of by the tokenizer.

(a) Hyphens: Identified to be of two types:-

- Without space on either sides, as in the case of compound nouns

eg. *mAtA-pitA(mother-father)*

(ii) With spaces, as in the case of

h85. *rAma ne kahA – yaha thIk hE*

'Ram' 'erg 'said' – 'this' 'proper' 'is'

(b) Quote marks (single and double both) : Identified to be of two types:-

1. opening
2. closing

9. Annotation Procedure

To maintain consistency in the data format and the annotation, it was decided to use 'Sanchay', a facility developed at IIIT, Hyderabad for the annotation task.

3.2.10. Conclusion

The annotation standards for POS tagging and chunking for Indian languages include 26 tags for POS (Table-1 in Appendix) and 11 chunk tags (Table-2 in Appendix. The tags are decided on coarse linguistic information with an idea to expand it to finer knowledge if required.

3.3. POS Tag Set for Indian Languages (Nov 2006, IIIT Hyderabad)

Sl No.	Category	Tag name	Example
1.1	Noun	NN	

1.2	NLoc	NST	
2.	Proper Noun	NNP	
3.1	Pronoun	PRP	
3.2	Demonstrative	DEM	
4	Verb-finite	VM	
5	Verb Aux	VAUX	
6	Adjective	JJ	
7	Adverb	RB	*Only manner adverb
8	Post position	PSP	
9	Particles	RP	bhI, to, hI, jI, hA.N, na,
10	Conjuncts	CC	bole (Bangla)
11	Question Words	WQ	
12.1	Quantifiers	QF	bahut, tho.DA, kam (Hindi)
12.2	Cardinal	QC	
12.3	Ordinal	QO	
12.4	Classifier	CL	
13	Intensifier	INTF	
14	Interjection	INJ	
15	Negation	NEG	
16	Quotative	UT	

17	Sym	SYM	ani (Telugu), endru (Tamil), bole/mAne (Bangla), mhaNaje (Marathi), mAne (Hindi)
18	Compounds	*C	
19	Reduplicative	RDP	
20	Echo	ECH	
21	Unknown	UNK	

It was decided that for foreign/unknown words that the POS tagger may give a tag ”UNK”

3.4.. Chunk Tag Set for Indian Languages

Sl. No	***Chunk Type***	***Tag Name***	***Example***
1	Noun Chunk	NP	*Hindi: ((merA nayA* **ghara))_NP** *“my new house”*
2.1	Finite Verb Chunk	VGF	*Hindi: mEMne ghara para khAnA ((khAyA*_VM*))*_**VGF**
2.2	Non-finite Verb Chunk	VGNF	*Hindi: mEMne* ***((khAte – khAte*_VM*))*_VGNF** *ghode ko dekhA*
2.3	Infinitival Verb Chunk	VGINF	*Bangla : bindu Borabela* ***((snAna karawe))*_VGINF** *BAlobAse*

Sl. No	***Chunk Type***	***Tag Name***	***Example***
2.4	Verb Chunk (Gerund)	VGNN	*Hindi: mujhe rAta meM ((**khAnA**_**VM**))_**VGNN** acchA lagatA hai*
3	Adjectival CHunk	JJP	*Hindi: vaha laDaZkI hE*((*suMdara*_**JJ** *sI*_RP))_**JJP**
4	Adverb Chunk	RBP	*Hindi : vaha* ((*dhIre-dhIre*_**RB**))_**RBP** *cala rahA thA*
5	Chunk for Negatives	NEGP	*Hindi:* **((*binA*))_NEGP ((***kucha***))**_NP **((*bole*))_VG ((***kAma***))**_NP ((*nahIM calatA*)**)_VG**
6	Conjuncts	CCP	Hindi: ((*rAma*))_NP ***((Ora))*_CCP** ((*SyAma*))_NP
7	Chunk Fragments	FRAGP	Hindi; ***rAma*** (*jo merA baDZA bhAI hE*) ***ne*** *kahA...*
8	Miscellaneous	BLK	

3.5. Tag set sample for Hindi

S.No.	English		Hindi	
1	**Noun**	***N***	*संज्ञा*	
	common	**N_NN**	**जातिवाचक**	काम, आम, शहर, जानवर etc

	Proper	**N_NNP**	*व्यक्तिवाचक*	बीजेपी, भाजपा
	Verbal		क्रियामूलक	
	Nloc	**N_NST**	देश-काल सापेक्ष	आगे, पीछे, ऊपर, नीचे, बाद, पहले, अंदर and बाहर
2	**Pronoun**	***PRP***	*सर्वनाम*	
	Personal	**PR_PRP**	*व्यक्तिवाचक*	मै, हम, मेरा, हमारा, मुझे, हमे, मुझी, हमीं, तुम, तुमहारा, तुमहे, तुझी
	Reflexive	**PR_PRF**	निजवाचक	अपना/अपने/अपनी, सवयं and खुद.
	Reciprocal	**PR_PRC**	पारस्परिक	आपस मे, परसपर
	Relative	**PR_PRL**	संबंध- वाचक	जो and its inflected forms
	Wh-words	**PR_PRQ**	प्रश्नवाचक	कब, कया, कौन, कहा, कैसे etc.
3	**Demonstrative**	**DM**	**निश्चयवाचक/संकेतवाचक**	
	Deictic	**DM_DMD**		यह, वह, ये, and वे.
	Relative	**DM_DMR**	सम्बन्ध वाचक	जो, जस
	Wh-words	**DM_DMQ**	प्रश्नवाचक	कब, कया, कौन, कहा, कैसे etc.

4	**Verb**	***V***	*क्रिया*	
	Auxiliary Verb	**V_VAUX**	सहायक क्रिया	रहा, होगा in पढ़ रहा होगा
	Main Verb	**V_VM**	मुख्य क्रिया	पढ़ in पढ़ रहा होगा
	Finite	**V_VM_VF**	परिमित	
	Infinitive	**V_VM_VNF**	अनंत	
	Gerund	**V_VM_VNG**	क्रियावाचक	
	Non-Finite	**V_VM_VNF**	गैर-परिमित	
5	**Adjective**	**JJ**	*विशेषण*	बड़ा, छोटा, लाल, सुंदर, चर, अचर, अिगरम etc.
6	**Adverb**	**RB**	*क्रिया-विशेषण*	धीरे, जलदी, तेज, etc.
7	**Post Position**	**PSP**	परसर्ग	मे, का, के, की, ने, पर, etc.
8	**Conjunction**		योजक	
	Co-ordinator	**CCD**	समन्वयक	और
	Subordinator	**CCS**		अगर, कयोिक, तो, िक, चूंिक etc.
	Quotative		उक्ति-वाचक	
9	**Particles**	**RP**	अव्यय	
	Default	**RP_RPD**	व्यतिक्रम	ही, तो and भी.
	Classifier		वर्गीकारक	

	Interjection	**RP_INJ**	विस्मयादिबोधक	ओह, आह, हाय, उफ, अरे, हे, ओ, अबे, etc.
	Negation	**RP_NEG**	नकारात्मक	नहीं, न, ना, मत, िबना etc.
	Intensifier	**RP_INTF**	तीव्रक	बेहद, बहुत, अतयंत, etc.
10	**Quantifiers**	**QT**	संख्यावाची	
	General	**QT_QTF**	सामान्य	थोड़ा, बहुत, जयादा, कम, कुछ
	Cardinals	**QT_QTC**	गणनासूचक	1, 2, 3, एक, दो, तीन etc.
	Ordinals	**QT_QTO**	क्रमसूचक	पहला, दसू रा, तीसरा etc. पाचवीं, बीसवा, तीसवे etc.
11	**Residuals**	**RD**	अवशेष	
	Foreign word	**RD_RDF**	विदेशी शब्द	
	Symbol	**RD_SYM**	प्रतीक	*, @, #, $, % etc.
	Unknown	**RD_UNK**	अज्ञात	
	Punctuation	**RD_PUNC**	विरामादि-चिह्न	
	Echowords	**RD_ECH**	प्रतिध्वनि-शब्द	पानी-वानी, खाना-वाना etc.

3.6. Linguistic Standards for Indian Language (IL) Corpora Annotation POS Categories and Labels (June 12, 2010)

POS for Telugu

Sl. No	Category			Label	Annotation Convention**	Examples	Remarks
	Top level	Subtype (level 1)	Subtype (level 2)				
1	**Noun**			**N**	**N**	abbAyi, puswakaM,	
1.1		Common		NN	N__NN	abbAyi, puswakaM	
1.2		Proper		NNP	N__NNP	kqRNudu, rAvu, reVddi	
~~1.3~~		~~Verbal~~		~~NNV~~	~~N__NNV~~	NA	Not Required
1.4		Nloc		NST	N__NST	munupu, kiMxata, moVxata, akkada	
2	**Pronoun**			**PR**	**PR**	wanu, vAdu, axi	
2.1		Personal		PRP	PR__PRP	wanu, nenu, nuvvu	
2.2		Reflexive		PRF	PR__PRF	svayaM	
2.3		Relative		PRL	PR__PRL	eVvadu, exi eVppudu, eVkkada,	
2.4		Reciprocal		PRC	PR__PRC	parasparaM	
2.5		Wh-word		PRQ	PR__PRQ	eVvadu, exi eVppudu, eVkkada,	

3	**Demonstrative**			**DM**	**DM**	A, I, e,	
3.1		Deictic		DMD	DM__DMD	A, I	
3.2		Relative		DMR	DM__DMR	e	
3.3		Wh-word		DMQ	DM__DMQ	e	
4	**Verb**			**V**	**V**	ammAnu, koVttanu, le, powAnu, rA,padiMxi, posi, powe	
4.1		Main		VM	V__VM	ammAnu, koVttanu, le, powAnu, rA,padiMxi, posi, powe	
4.1.1			Finite	VF	V__VM__VF	ammAnu, koVttanu, le, powAnu, rA,padiMxi,	
4.1.2			Non-finite	VNF	V__VM__VNF	wini, wine winaka, wiMtU	
4.1.3			Infinitive	VINF	V__VM__VINF	wina, vina, ceVppa	
4.1.4			Gerund	VNG	V__VM__VNG	winadaM, vinadaM	
4.2		Auxiliary		VAUX	V__VAUX	uMxi, kAxu, lexu, padu	
5	**Adjective**			**JJ**		cakkani, maMci, peVxxa	
6	**Adverb**			**RB**		veVMt ane, wvaragA	
7	**Postposition**			**PSP**		ku, cewa, lona,bayata	
8	**Conjunction**			**CC**	**CC**	mariyu, iMkA , ani	

						Ewe, kAnI	
8.1		Co-ordinator		CCD	CC__CCD	mariyu, iMkA , kAnI	
8.2		Subordinator		CCS	CC__CCS	ani, aMte, aMtU	
8.2.1			Quotative	UT	CC__CCS__UT	ani,	vIdu vAde ani weVlisiMxi,
8.2.2			Introducer or Identifier		CC__CCS__ID	aMte, aMtU	vIdu aMte nIku weVlusA? AkASaM aMtU oVka paxArWaM lexu.
9	**Particles**			**RP**	**RP**	ayiwe,kUdA , mAwraM	
9.1		Default		RPD	RP__RPD	ayiwe,kUdA , mAwraM	
9.2		Classifier		CL	RP__CL	maMxi, guru	
9.3		Interjection		INJ	RP__INJ	are, ayyo, oy,ammo	
9.4		Intensifier		INTF	RP__INTF	cAla, eVnni, eVMwo	
9.5		Negation		NEG	RP__NEG	wappa,vinA	
10	**Quantifiers**			**QT**	**QT**	koVxxi, cAlA, oVka, koVMceVM	
10.1		General		QTF	QT__QTF	koVxxi, cAlA, koVMceVM	
10.2		Cardinals		QTC	QT__QTC	oVkati, reVMdu, mUdu	

10.3		Ordinals		QTO	QT__QTO	oVkato, reVMdava, mUdava	
11	**Residuals**			**RD**	**RD**		
11.1		Foreign word		RDF	RD__RDF		A word written in a script other than the script of the current text
11.2		Symbol		SYM	RD__SYM	$, &, *, (,)	such as $, &
11.3		Punctuation		PUNC	RD__PUNC	.,,, :, ;	Punctuations markers
11.4		Unknown		UNK	RD__UNK		
11.5		Echowords		ECH	RD__ECH	(illu-) gillu, (puli-) gili	

** The annotation is to be done using the lowest level tag of the type hierarchy. Once the lower level tag is selected, the higher level tags should be stored automatically.

3.7. Parts of Speech (PoS) Annotation Guideline for Hindi Following the BIS Tagset (Draft)

The Bureau of Indian Standards (BIS) Tagset has recommended the use a common tagset for the part of speech annotation of Indian languages. The tagset, incorporating the advice of the experts and the stakeholders in the area of natural language processing and language technology of Indian languages, has to be followed in the annotation tasks taking place in Indian languages after August, 2010. Although no formal guideline has been presented so far, the annotations taking place under the Indian Languages Corpora Initiative (ILCI) program is following the BIS tagset as proliferated. This document is an attempt to present a guideline for annotating Hindi using the BIS tagset. The BIS tagset has a total of 38 annotation level tags which are common to all the Indian languages covered under this tagset. Out of these 38 tags, a total of 30 tags have been recommended to be applied for Hindi. Parts of speech have been divided into a total of 11 top level categories which are further divided into 30 subcategories. The summary of the Hindi tagset presented in Table 1 below.

Table1. Summary of the BIS Tagset for Hindi

Category Subcategory

Noun Common Proper Nloc

Noun Labels : **N_NN N_NNP N_NST**

Pronoun Personal Reflexive Relative Reciprocal Wh-word

Pronominal Labels: **PR_PRP PR_PRF PR_PRL PR_PRC PR_PRQ**

Demonstrative

Deictic Relative Wh-word

Demonstrative Labels: **DM_DMD DM_DMR DM_DMQ**

Verb Main Auxiliary

Verbal Labels: **V_VM V_VAUX**

Adjective Adjective

Adjectival Label: **JJ**

Adverb Adverb

Adverbial Label: **RB**

Postposition Postposition

Postpositional Label: **PSP**

Conjunction Co-ordinator Subordinator

Conjunctival Labels: **CCD CCS**

Particles Default Interjection Intensifier Negation

Particle Labels: **RP_RPD RP_INJ RP_INTF RP_NEG**

Quantifiers General Cardinals Ordinals

Quantifier Labels: **QT_QTF QT_QTC QT_QTO**

Residuals Foreign word Symbol Punctuation Unknown Echo-words

Residual Labels: RD_RDF RD_SYM RD_PUNC RD_UNK RD_ECH

Table 1 above shows all the eleven categories and their respective sub-categories followed by the partof-speech labels for the subcategories. Except for the three categories of adjective, adverb and postposition, all the categories have some two or more sub-categories. The category of residual is though not part of the language, it is part of the text which is to be annotated. Therefore, this category

also has extra-linguistic elements appearing in the text sub-categorized.

1. Syntactic Function vs Lexical Category

A word belonging to a particular lexical category may function differently in a given context. For example, the lexical category of हिरजन in Hindi is a noun. However, functionally, हिरजन is used as an adjective in ex.1 below,

Ex.1. एक िदन पाँच बजे खबर आयी िक कोई हिरजन बालक उनसे िमलनाचाहता हैO

One day five o'clock news came that some harijan boy him meet wants is One day, a message came at five o'clock that a 'harijana' boy wanted to meet him.

Such cases require a decision on whether to tag a word according to its lexical category or by its syntactic category. Since the word in a context has syntactic relevance, it appears natural to tag it based

on its syntactic information. However, such a decision may lead to further complications. In annotation, the syntactic function of a word is not considered for POS tagging. Since the word is always tagged according to its lexical category there is consistency in tagging. This reduces confusion involved in manual tagging. Also the machine is able to establish a word-tag relation which leads to efficient machine learning. In short, it is recommended that syntactic and semantic/pragmatic functions were not to be the basis of deciding a POS tag.

2. Nouns (NN)

The top level category of Noun has been divided into three sub-categories- common, proper and spatiotemporal. While the common and proper nouns are pretty easy to decipher, the category of spatiotemporal nouns are a bit tricky. The words such as ऊपर, नीचे, आगे, पीछे can be used in several ways. The most typical is their use postposition in conjunction with another postpositional word के as in के ऊपर, के नीचे, के पीछे etc.

2.1. Common Nouns (NN_NNC)

The nouns that simply function as noun and are content words should be marked as the common noun.

This includes the general variety of all the nouns, e.g. काम, आम, शहर, जानवर etc.

2.2.Proper Nouns (NN_NNP)

The proper nouns are basically some specific names which denote to one particular entity. It includes

the names of person, place or thing. The examples would be राम, मोहन, कोलकाता, िदलली, िहमालय, कोका-कोला etc.

The names of months, days, seasons are not to be marked as NSTs even though they denote time. They

are in fact common nouns.

सावन/N_NNC के महीने मे बािरश होती है।

मै यहा सात जनवरी/N_NNC को आया था।

No separate tag has been assigned to abbreviations. They are to be marked as proper nouns. Acronyms,

if used as proper nouns, should be marked as proper nouns and if common then as common.

बीजेपी/N_NNP, भाजपा/N_NNP,

2.3. Spatio-Temporal Nouns (NN_NST)

There is a specific set of words which function sometimes as argument of verb and sometimes as the postposition. The sub-category of NST has been included to specifically capture these words. In these

words are as follows:

There is a specific set of words which function sometimes as argument of verb and sometimes as the postposition. The sub-category of NST in the standards document has been included to specifically capture these words. In these words are as follows:

आगे, पीछे, ऊपर, नीचे, बाद, पहले, अंदर and बाहर

So, irrespective of their syntactic function, these words have to be constantly marked as the NST.

वह आगे/NN_NST जा रहा था

घर के आगे/NN_NST एक पेड़ था।

तुम मेरे बाद/NN_NST आए थे।

Here is the explanation (with some modifications to suit here) that appears in the standards document

for the conclusion drawn above for the category of NST:

आगे *Age* (front), पीछे *plche*(behind), ऊपर *Upara*(above/upstairs), नीचे *nlce*(below/down), बाद *bAda*(after),पहले *pahale*(before), अंदर *andara*(inside), बाहर *bAhara (outside)* etc.

The tag N_NST has been included to cover an important phenomenon of Indian languages. Certain expressions such as ऊपर, नीचे, पहले, आगे etc. are content words denoting time and space. These expressions, however, are used in various ways. These words often occur as temporal or spatial arguments of a verb in a given sentence taking the appropriate *vibhakti* (case marker):

Ex2. वह ऊपर सो रहा था

*vaha **Upara** so rahA thA* .

'he' 'upstairs' 'sleep' 'PROG' 'was'

“He was sleeping upstairs”.

Ex3. वह पहले से कमरे मे बैठा था

*vaha **pahale se** kamare meM bEThA thA*.

'he' 'beforehand' 'from' ' room' 'in' 'sitting' 'was'

“He was sitting in the room from beforehand”

Ex4. तुम बाहर बैठो

*tuma **bAhara** bETho*

'you' 'outside' 'sit'

“You sit outside”.

Apart from functioning like an argument of a verb, these elements also modify another noun taking postposition 'kA'.

Ex5. उसका बड़ा भाई ऊपर के िहससे मे रहता है

*usakA baDZA bhAI **Upara ke hisse meM** rahatA hE*

'his' 'elder' 'brother' 'upstairs' 'of' 'portion' 'in' 'live' 'PRES'

“His elder brother lives in the upper portion of the house”.

Apart from occurring as a nominal expression, they also occur as a part of a postposition along with 'ke'.

For example,

Ex6. घड़े के ऊपर थाली रखी है

*ghaDZe **ke Upara** thAlI rakhI hE.*

'pot' 'of' 'above' 'plate' 'kept' 'is'

The plate is kept on the pot”.

Ex7. तुम घर के बाहर बैठो

*tuma ghara ke **bAhara** bETho*

'you' 'home' 'of' 'outside' 'sit'

“You sit outside the house”.

'Upara' and '*bAhara*' are parts of complex postpositions *'ke Upara'* and '*ke bAhara*' in (ex5) and (ex6)

respectively which can be translated into English prepositions 'on' and 'outside'.

For tagging such words, one possible option is to tag them according to their syntactic function in the given context. For example in (ex6) above, the word '*Upara*' is occurring as part of a postposition or a relation marker. It can, therefore, be marked as a postposition. Similarly, in (ex2) and (ex5) above, it is a

noun, therefore, mark it as a noun and so on. Alternatively, since these words are more like nouns, as is

evident from above they can be tagged as nouns in all their occurrences. The same would apply to '*bAhAra'* (outside) in examples (ex3), (ex4) and (ex7). However, if we follow any of the above approaches we miss out on the fact that this class of words is slightly different from other nouns. These are nouns which indicate 'location' or 'time'. At the same time, they also function as postpositions in certain contexts. Moreover, such words, if tagged according to their syntactic function, will hamper machine learning. Considering their special status, it was considered whether to introduce a new tag, NST, for such expressions. The following five possibilities were discussed:

a) Tag both (ex4) & (ex7) as NN

b) Tag both (ex4) & (ex7) as NST

c) Tag (ex4) as NN & (ex7) as NST

d) Tag (ex4) as NST & (ex7) as PSP

e) Tag (ex4) as NN & (ex7) as PSP

After considering all the above, the decision was taken in favour of (b). The decision was primarily based on the following observations:

(i) *'bAhara'* in both (ex4) and (ex8) denotes the same expression (place expression 'outside')

(ii) In both (ex4) and (ex7), *'bAhara'* can take a vibhakti like a noun (**bAhara ko** bETho, ghara **ke bAhara ko** bETho)

(iii) If a single tag is kept for both the usages, the decision making for annotators would also be easier.

Therefore, a new tag **NST** is introduced for such expressions. The tag **NST** will be used for a finite set of such words in any language. For example, Hindi has आगे *Age* (front), पीछे

pIche(behind), ऊपर *Upara*(above/upstairs), नीचे *nIce*(below/down), बाद *bAda*(after),पहले *pahale*(before), अंदर *andara*(inside), बाहर *bAhara (outside)* etc.

3. Pronouns (PR)

The category of pronoun has been divided into five sub-categories. These include personal, reflexive,

relative, reciprocal and wh-word. These categories should be self-explanatory and follows the same

definitions as posited in common linguistic literature.

3.1 Personal Pronouns (PR_PRP)

Personal pronouns cover all the pronouns that denote to person, place or thing. This includes the all

their cases as well for example: मै, हम, मेरा, हमारा, मुझे, हमे, मुझी, हमीं, तुम, तुमहारा, तुमहे, तुझी etc.

3.2 Reflexive Pronoun (PR_PRF)

Reflexive pronouns are the ones that denote to ownership to its antecedent which can be either a noun

or a pronoun. The only examples of reflexive pronouns in Hindi are the अपना/अपने/अपनी, सवय ं and खुद_____.

(Please report to me if you find any other word that can go into this category) E.g.:

वह अपना/PR_PRF खाना खुद/PR_PRF पकाता है।

राम सवयं/PR_PRF सामथयरवान है।

वह अपने/PR_PRF गाव गया है।

3.3 Relative Pronouns (PR_PRL)

The relative pronouns are those pronouns whose antecedent can be either a noun or a pronoun. However, these pronouns do not make any difference in number or gender as in the case of personal

pronouns. The relative pronoun in Hindi is represented by जो and its inflected forms e.g.

जो/PR_PRL करता है वो भरता है।

वह लड़का जो/PR_PRL आया था चला गया।

िजसे/PR_PRL आपने बुलाया था वह नहीं आया।

The list of relative pronouns may be exhaustive. One has to make sure that one makes the appropriate

difference between the relative pronoun and the relative demonstrative pronoun. Note that in the following sentences the occurrences of जो and िजस are demonstrative relatives (DM_DMR) and not relative pronouns (PR_PRL)

जो/DM_DMR लड़का यहा आया था वह चला गया।

िजस/DM_DMR लड़के को आपने बुलाया था वह नहीं आया।

3.4 Reciprocal Pronoun (PR_PRC)

Reciprocal pronouns denote some reciprocity. This is commonly denoted by आपस मे, परसपर

वे लोग आपस/PR_PRC मे झगड़ा कर रहे थे।

परसपर/PR_PRC बात करके सुलझा लीिजए।

3.5 Wh-word Pronouns (PR_PRQ)

The wh-word pronouns are typically the pronouns that are used to ask questions. These words are कब,

कया, कौन, कहा, कैसे etc.

कया/PR_PRQ लाए हो?

कया/PR_PRQ तुमने खा िलया?

िफर कब/PR_PRQ आओगे?

4. Demonstratives (DM)

The category of demonstrative has been separated from the category of pronouns as the demonstratives mainly indicate about a noun and does not act as anaphora. The demonstratives have been sub-categorized into three divisions- deictic, relative and wh-words.

4.1.Deictic (DM_DMD)

The deictic demonstratives are default demonstratives that demonstrate the noun it modifies. The deictic demonstratives in Hindi are typically यह, वह, ये, and वे. These always occur before the noun they modify.

यह/DM_DMD शहर बहुत पराचीन है।

उस/DM_DMD घर की छत पकी है।

4.2.Relative Demonstrative (DM_DMR)

The relative demonstrative occur in the same form as the relative pronoun. The difference is only that

these relatives are always followed by a noun that it modifies.

िजस/DM_DMR गाव मे मै गया था वह बहुत सुंदर था।

जो//DM_DMR नहर टूट गई थी उसकी मरममत की जा रही है।

4.3.Wh-Word Demonstrative (DM_DMQ)

The wh-demonstratives are the same wh-words (or question words) which act as wh-pronouns. The difference is that in their demonstrative function they do not ask question, rather only demonstrates.

The wh-word demonstratives are कोई, ि◌कसी, कौन etc.

कोई/DM_DMQ लड़का आया था।

यह बात ि◌कसी/DM_DMQ से छुपी नहीं है।

जाने कौन/DM_DMQ यहा रहता है।

जाने कहा/DM_DMQ से आई है।

5. Verb (V)

The verb in the BIS has been divided into two categories- main and auxiliary. The dual distinction is to decipher main verb and the auxiliary verb(s) in any sentence. While the auxiliary verb is a closed set of verb, the main verb can be anything from a root verb to any of its inflected forms. Each sentence or clause must have a main verb. A sentence can have one more auxiliary verbs. As in the following sentences:

Simple Verbal occurrences

मै/PR_PRP मोहन/N_NNP हूं/V_VM ।/RD_PUNC

वह/PR_PRP पढ़ता/V_VM है/V_VAUX ।/RD_PUNC

वह/PR_PRP पढ़/V_VM रहा/V_VAUX है/V_VAUX ।/RD_PUNC

वह/PR_PRP पढ़/V_VM रहा/V_VAUX होगा/V_VAUX ।/RD_PUNC

Modal Auxiliaries

उसे/PR_PRP अब/N_NNC तक/PSP आ/V_VM जाना/V_VAUX चाि◌हए/V_VAUX था/V_VAUX ।/RD_PUNC

Compound Verbs

मै/PR_PRP अब/RB सवयं/PR_PRF समझ/V_VM सकता/V_VAUX हूं/V_VAUX ।/RD_PUNC

उसने/PR_PRP ◌िकताब/N_NNC फाड़/V_VM दी/V_VAUX है/V_VAUX ।/RD_PUNC

उसने/PR_PRP मुझे/PR_PRP नहीं/RP_NEG जाने/V_VM ◌िदया/V_VAUX ।/RD_PUNC

Conjunct Verbs

The conjunct verbs are preceded by a noun or an adjective. This preceding noun or adjective are marked with their own category while the verb (which is usually a set of verbs called sometimes light verb or vector verbs) coming after, it will be marked as the main verb.

वह/PR_PRP गुससे/N_NNC से/PSP लाल/JJ हो/V_VM रहा/V_VAUX था/V_VAUX ।/RD_PUNC

उसने/PR_PRP तुमहे/PR_PRP मूखर/JJ बना/V_VM ◌िदया/V_VAUX ।/RD_PUNC

Verbal Nouns/ Gerunds

In the BIS standards document, the category of verbal nouns or gerund is not considered for Hindi o other Indo-Aryan languages. Such verbs function as the noun in the sentences but are capable of taking

their own arguments. For this very reason, they are to be marked as the verbs. Even though they do not

function as the main verb in the sentence, they are to be marked with the label of main verb.

खेलना/V_VM सवासथय/N_NNC के/PSP ◌िलए/PSP अचछा/JJ होता/V_VM है/V_VAUX ।/RD_PUNC

हंसने/N_NNC से/PSP अवसाद/N_NNC कम/JJ होता/V_VM है/V_VAUX ।/RD_PUNC

मै/PR_PRP आपसे/PR_PRP ◌िमलने/V_VM को/PSP उतसुक/JJ हूं/V_VM ।/RD_PUNC

Nouns Derived from Verbs

Nouns derived from verbs (such as पढ़ाई, खुदाई, पुताई, etc.) are to be marked as nouns. These are marked as nouns because they cannot take an argument of their own in the sentence.

सुबह/N_NN की/PSP पढ़ाई/N_NN याद/N_NN रहती/V_VM है/V_VAUX ।/RD_PUNCT

*सुबह की िकताब पढ़ाई याद रहती है।

खुदाई/N_NN का/PSP काम/N_NN चल/V_VM रहा/V_VAUX है/V_VAUX ।/RD_PUNCT

*गढढा खुदाई का काम चल रहा है।

Participial constructions of verbs acting as modifiers

The standards document does not specify anything about the participial and the –कर constructions which also abound in Hindi. I presume that these would also be taken as main verbs.

दौड़ता/V_VM हुआ/V_VAUX लड़का/N_NN िगर/V_VM पड़ा/V-VAUX ।/RD_PUNCT

सैिनक/N_NN दौड़कर/V_VM आया/V_VM ।/RD_PUNCT

दौड़ने/V_VM वाला/V_VAUX लड़का/N_NN परथम/QT_QTO आया/V_VM ।/RD_PUNCT

6. Adjective (JJ)

Adjective has not been sub-divided into any categories. There is one category for an adjective which is self-explanatory. These are mostly attributive adjectives. For quantifiers, there is a separate category defined. Examples: बड़ा, छोटा, लाल, सुंदर, चर, अचर, अिगरम etc.

7. Adverb (RB)

Adverb also is mono-category part-of-speech. The standards document says that the category of adverb

(RB) is only for manner adverbs. For example, words like धीरे, जलदी, तेज, etc. are adverbs. However, there are many other words which qualify for adverbial category, but are not given this

category. So, some confusion remains. For the example the words like िबलकुल, एकदम, हमेशा, इसीिलए, and अकसर are such words which mostly modify the verb or the whole of a clause and are not manner adverbs. For such words, it has been advised that their category should be examined closely and the label should be given accordingly. For example, the word इसीिलए/इसिलए has been suggested to be tagged as pronominal pronoun because it is made up of "इस/इसी के िलए" and because in this the word 'इस/इसी' is the headword which is a pronoun. Thereby, the category for the word "इसिलए/इसीिलए" should be given as pronoun. However, even if we concede for this, what about the words like अत: which is just another word for the same purpose as that of "इसिलए/इसीिलए" Considering the use of words such as िबलकुल, एकदम, हमेशा, अकसर, अतः, we need to reconsider whether we should include a category of adverbs other than the manner adverbs to address these words.

8. Postposition (PSP)

Postpositions are all the parts-of-speech that work as case marker. Words like मे, का, के, की, ने, पर, etc. are examples of postposition.

राम का/PSP भाई सुंदर है।

गाव मे/PSP लोग कुए ंका/PSP पानी पीते है।

There are postpositions in Hindi and other Indian languages that are formed with more than words contributing to it. Such postpositions are called complex postpositions. Examples of such postpositions

are के िलए, के ऊपर, के बाद, के पीछे etc. Such complex postpositions are to be marked as PSP for all the

words that constitute it. Example:

मैने राम के/PSP िलए/PSP यह पुसतक लाई है।

मेज के/PSP ऊपर/PSP पुसतक रखी है।

मरे े बाद/PSP राम आएगा।

9. Conjunction (CC)

Conjunctions are non-content words that act as joiners of phrases or clauses within a sentence. The category of conjunction has been divided into two sub-categories of coordinator and subordinator.

9.1. Coordinators (CC_CCD)

Coordinators are typically the words that join two phrases, of the same category. It can join one or more

of a noun phrase, a verb phrase or a clause. The common examples are और, तथा, व, बिलक, िकनतु, परनतु etc.

राम और/CC_CCD शयाम घर जा रहे है।

उसने ऐसा केवल कहा ही नहीं बिलक/CC_CCD िकया भी।

9.2. Subordinators (CC_CCS)

Subordinator conjunctions typically conjoin two clauses and the second clause is subordinated. That is the clause conjoined by the subordinator word is the subordinate clause against the main clause. Typical

subordinator words in Hindi are अगर, कयोिक, तो, िक, चूंिक etc.

अगर/CC_CCS आज वह आया तो/CC_CCS मै उसे बताऊं गा।

उसने कहा िक/CC_CCS वह आज नहीं आएगा।

10. Particles

Particles are words that do not decline and also do not fall into any other categories described above and elsewhere. Four sub-categories - default, interjection, intensifier and negation - have been created

to cover the particles in Hindi.

10.1. Default

The default particles are ही, तो and भी.

वह आने ही/RP_RPD वाला है।

वह पढ़ ही तो/RP_RPD रहा था।

जाने भी/RP_RPD दो अब।

10.2. Interjection

Interjections are particles which denote exclamation utterances. The common exclamatory marks in Hindi are ओह, आह, हाय, उफ, अरे, हे, ओ, अबे, etc.

ओह/RP_INJ ये कया हो गया!

उफ//RP_INJ हद हो गई यार!

10.3. Intensifier

Intensifiers are words that intensify the adjectives or adverbs. The common intensifiers in Hindi are बेहद,

बहुत, अतयंत, etc.

वह बेहद/RP_INT कमजोर हो चला है।

वह बहुत/RP_INT ही धीरे चलता है।

यह अतयतं /RP_INT मा◌िमरक कहानी है।

10.4. Negation

The negation particles are the words that indicate negation. These include नहीं, न, ना, मत, िबना etc.

आज वह नहीं/RP_NEG आया।

शोर ना/RP_NEG करना।

11. Quantifiers QT)

Quantifiers are the words that modify nouns or adjectives and indicate quantity. These have been subcategorized into three parts- general, cardinals and ordinals.

11.1. General (QT_QTF)

The general quantifiers do not indicate any precise quantity, e.g. थोड़ा, बहुत, जयादा, कम, कुछ, etc.

11.2. Cardinals (QT_QTC)

The cardinal quantifiers are absolute numbers, either in digits or in words such as 1, 2, 3, एक, दो, तीन etc.

11.3. Ordinals (QT_QTO)

The ordinals denote the order part of the digits such as पहला, दसू रा, तीसरा etc. The ordinals in Hindi also inflect for gender and number and take the oblique case marker, thus पाचवीं, बीसवा, तीसवे etc.

12. Residuals (RD)

The category of residuals has been demarcated for the words that are usually not intrinsic part of the language/speech. Divided into five parts, these include foreign words, symbols, punctuations, unknown

words and echo-words.

12.1. Foreign Words (RD_RDF)

The foreign words are all the words that are not written in the Devanagari script.

12.2. Symbols (RD_SYM)

The symbols are the characters that are not part of the regular Devanagari script such as *, @, #, $, %

etc.

12.3. Punctuations (RD_PUNC)

Punctuations include the characters that are considered as the regular punctuation marks in Hindi, e.g.

(,),,,!,?,- etc.

12.4. Unknown (RD_UNK)

Unknown words would the words for which a category cannot be decided by the annotator. These may

include words from phrases or sentences from a foreign language written in Devanagari.

12.5. Echo-Words(RD_ECH)

The echo-words are the words that are formed by the morphological process known as echo-formation

e.g. पानी-वानी, खाना-वाना etc.

3.8. Comments for Tamil POS tags

1. Tamil does not have a relative pronoun. The examples yaar, etu, eppootu, enkee given under Relative Pronoun (2.3) are not relative pronouns. They are only Wh-words.
2. Also there is no relative demonstrative in Tamil. The example enta given for Relative Demonstrative (3.2) is only a wh-word.

3. The examples given for Gerund and verbal noun are wrong. Though Gerund and Verbal nouns are nouns, separate POS tag symbols are assigned to them in order to distinguish them from ordinary nouns as they retain the S structure of verbs. For example, in the sentence "avan tinamum aappiL caappiTuvatu uNTu", caappiTuvatu retains the argument structure of the verb capiTu. Similarly, for example, in the sentence "kaalaiyil kuLirnta niiril kuLittal nallatu", the verbal noun kuLittal retain the argument structure of the verb kuLi. That is the reason in the POS tag list Gerunds and Verbal nouns, though nominal in character, are distinguished form ordinary nouns by assigning them separate tag symbols. But the derived nouns paTippu, naTattai, naTai are simply nouns and they do not retain the argument structure of the concerned verbs. So they need not be given separate tag label. They can be taken as ordinary nouns. So the examples given for Gerund tag (4.1.4) are wrong. They should be vizuvatu, cirippatu, tuunkuvatu. On the same ground the examples given under Verbal tag (4.2) are also wrong. They should be paTittal, naTattal, ceytal.
4. The transliteration of examples given uder particle (9) maTTUm is to be corrected as maTTum. Similarly the example given under 9.1 maTTUm should be corrected as maTTum.
5. The transliteration of the example under 9.3 teey should corrected as Teey.

3.9.Comments for Malayalam POS tags:

1. The transliteration of the examples given for tags is not uniform. Long vowels are given capital letters mostly, but some times doubling of the letters.
2. The distinction between the alveolar l and retroflex L are not mainted properly.
3. Similarly distinction between dental stop t and the retroflex stop T are also not maintained properly.
4. The following list gives the wrong and correct forms.

Wrong ones	Corrected ones

1.	vItu	vITu	
1.1.	vItu	vITu	
	vellam	veLLam	
	pattam	paTTam	
2.	aval	avaL	
2.1.	naan	njaan	
3.3.	ennane	engngane	
4.	Annu	ANu	(Malayalam word is also written wrongly) written)
4.1.	cirri	ciri	(Malayalam word is missing)
	Annu	ANu	(Malayalam word is also written wrongly)
4.1.3.	pOkku	pOkuka	(Malayalam word is also written wrongly)
	kazhikkee		(Malayalam word is written wrongly)
4.3	kAnuka	kANuka	
5.	azakulla	azhakuLLa	
6.	kUtutal	kUTutal	
7.	kUte	kUTe	
8.	ennalum	ennAlum	
	enkilum	engkilum	
9.	kute	kUTe	
9.4.	valare	vaLare	
10.	kuracchu	kuRaccu	

	niraccu	niRaccu
	dharalam	dhArALam
10.1.	kuraccu	kuRaccu
	niraccu	niRaccu
	dharalam	dhArALam
10.2.	rantu	raNTu
10.3.	rantam	raNTAm

5. In 2.3. ArO is given as relative pronoun. There is no relative pronoun for Malayalam.
6. In 3.2. Etu is given as relative demonstrative. There is no relative demonstrative for Malayalam.
7. In 4.2. paTittam, naTattam, naTanam are given as verbal noun. As argued under Tamil these nominal forms from verbs behave like ordinary noun forms. They need not be given special status here.
8. There is no gerund listed. The gerund forms need to be dealt separately as they keep the argument structure of the concerned verb intact. So there should be a row for gerund (4.1.4). The examples like paTiccatu, paTikkunnatu, naTannatu, naTakkunnatu have to be given as gerund.
9. In 9.2. peer is given as classifier. There is no classifier for Malayalam.
10. In 9.5. illa, alla are given as negation particle. But illa and alla are negative auxiliary verbs and they are not negative particles.

3.10. PARTS OF SPEECH TAGGER AND CHUNKER FOR MALAYALAM

Parts of Speech Tagger (POS) is the task of assigning to each word of a text the proper POS tag in its context of appearance in sentences. The Chunking is the process of identifying and

assigning different types of phrases in sentences. In this paper, a statistical approach with the Hidden Markov Model following the Viterbi algorithm is described. The corpus both tagged and untagged used for training and testing the system is in the Unicode UTF-8 format.

Introduction

Parts of Speech Tagging, a grammatical tagging, is a process of marking the words in a text as corresponding to a particular part of speech, based on its definition and context. this is the first step towards understanding any languages. It finds its major application in the speech and NLP like Speech Recognition, Speech Synthesis, Information retrieval etc. A lot of work has been done relating to this in NLP field.

Chunking is the task of identifying and then segmenting the text into a syntactically correlated word groups. Chunking has been traditionally defined as the process of forming groups of words based on local information. Chunking can be viewed as shallow parsing but does not specify their internal structure, or their role in the main sentence. This text chunking can be considered as the first step towards full parsing. Mostly Chunking occur after POS tagging. Chunking can be useful for information retrieval, information extraction, and question answering, since a complete chunk (Noun, Verb or Preposition Phrase) is likely to be semantically relevant for the requested information. This is very important for activities relating to Language processing.

3.11. STATISTICAL APPROACH TOWARDS TAGGING AND CHUNKING

The statistical methods are mainly based on the probability measures including the unigram, bigram, trigram and n-grams. The Hidden Markov Model is a finite set of states, each of which is associated with a (generally multidimensional) probability distribution. Transitions among the states are governed by a set of probabilities called transition probabilities. In a particular state an outcome or observation can be generated, according to the associated probability distribution. It is only the outcome, not the state visible to an external observer and therefore states are ``hidden" to the outside; hence the name Hidden Markov Model.

A Hidden Markov Model (HMM) is a statistical model in which the system modeled is thought to be a Markov process with the unknown parameters. In this model, the assumptions on which it works are the probability of the word in a sequence may depend on its immediate word presiding

it and both the observed and hidden words must be in a sequence. In a hidden Markov model, the state is not directly visible, but variables influenced by the state are visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. This model can represent the observable situations and in POS tagging and Chunking, the words can be seen themselves, but the tags cannot. So HMM are used as it allows observed words in input sentence and hidden tags to be build into a model, each of the hidden tag state produces a word in a sentence.

With HMM, Viterbi algorithm, a search algorithm is used for various lexical calculations. The Viterbi algorithm operates on a state machine assumption. It is a dynamic programming algorithm that is mainly used to find the most likely of the hidden states, results in a sequence of the observed words. This is one of the most common algorithms that implement the n-grams approach. This algorithm mainly work based on the number of assumptions it makes.

The algorithm assumes that both the observed and the hidden word must be in a sequence, which corresponds to the time. It also assumes that the sequence must be aligned. One of the basic views behind this algorithm is to compute most likely tag sequence occurring with the unambiguous tag until the correct tag is obtained. At each level most appropriate sequence and the probability including these are calculated. The Viterbi algorithm is a computationally simple method of finding the maximum a posteriori estimate of a sequence of observations.

In Malayalam, there are many chances where each word may come up with different tags. This is since Malayalam is morphologically rich and agglutinative language. According to the context also there are chances where the tags may be given differently.

3.12. TAGSET FOR POS TAGGING AND CHUNKING

Malayalam belongs to the Dravidian family of languages, inflectionally rich in the morphology. The words are formed mainly by adding of suffixes with the root or the stem word. Since words are formed by the suffix addition with root, the word can take the POS tag based on the root / stem. Hence it can be stated that the suffixes play major role in deciding the POS of the word.

Tag set for Parts of Speech Tagging

Main Tags	Representation

1	Noun	NN
2	Noun Location	NST
3	Proper Noun	NNP
4	Pronoun	PRP
5	Compound Words	XC
6	Demonstration	DEM
7	Post Position	PSP
8	Conjuncts	CC
9	Verb	VM
10	Adverb	RB
11	Particles	RP
12	Adjectives	JJ
13	Auxiliary Verb	VAUX
14	Negation	NEG
15	Quantifiers	QF
16	Cardinal	QC
17	Ordinal	QO
18	Question Words	WQ
19	Intensifiers	INTF
20	Interjection	INJ
21	Reduplication	RDP
22	Unknown Words	UNK
23	Symbol	SYM

For Chunking, mainly six tags are used. This is based on the grammatical or the syntactical category.

Tag set for Chunking

Main Tag set	Representation	
1	Noun Phrase Chunk	NP

2	Verb Finite Chunk	VGF
3	Non-Finite Verb Chunk	VGNF
4	Conjunction Chunk	CCP
5	Verb Chunk Gerund	VGNN
6	Negation Chunk	NEGP

3.13. SYSTEM TESTING AND RESULT

The system is trained using the manually tagged corpus. While training, the transition and e emission property matrix are calculated and the models of the language are build. The lexicon file created during the generation of the parameter contains the frequencies of the words and its tags that occurred in the training corpus. This is used for determining the lexical probability. The n-gram file that is also generated during the parameter generation contains the contextual frequencies for the unigrams, bigrams, and trigrams.

System Testing:

After training the system using the manually tagged corpus, the system can be now tested with the raw or untagged corpus. For the tagging pf the raw corpus, both the files which contains the modal parameter for the lexical and the contextual frequencies are required.

Result:

Following results were obtained while testing the raw corpus with in the system. The raw corpus used for testing was in Unicode. For training the system, ie for in the training phase, the tagger and chunker were trained with using about 15,245 tokens. This can be increased further to any extent there by increasing the accuracy of the system.

In case of Parts of Speech Tagging

Comparing 200 tokens, Overall result,

Equal : 181/200 (90.5%)

Different : 19/200 (9.5%)

In case of Chunking

Comparing 200 tokens, Overall result,

Equal : 184/200 (92.00%)

Different : 16/200 (8.00%)

For chunking, the system gives about 92% accuracy while for POS tagging it gave about 90.5% accuracy.

The Parts of Speech Tagging and Chunking for Malayalam using the statistical approach has been discussed. The system works fine with the Unicode data. The POS and Chunker were able tot assign tags to all the words in the test case. These also focus on the point that a statistical approach is also work well with highly morphologically and inflectionally rich languages like Malayalam.

4. COMPLEXITY IN POS TAGGING OF TAMIL

The paper aims to focuses on the Morphological complexity in Tamil language from the point of view of post tagging. The morphological complexity leads to the complexity of Tamil in POS tagging. Nouns get inflected for number and cases. Verbs get inflected for various inflections which include tense, finite and non-finite suffixes. Verbs are adjectivalized and adverbialized. Also verbs and adjectives are nominalized by means of certain nominalizers. Adjectives and adverbs do not inflect. So, many times we need to depend on syntactic function or context to decide upon whether one is a noun or adjective or adverb

PATS OF SPEECH IN TAMIL

The following parts of speech or word classes are identified for Tamil languages by modern grammarians:

1) Noun, 2) Verb, 3) Adjective, 4) Adverb, 5) Postposition, 6) Numeral, 7) Quantifier, 8) Words of conjunction, 9) Exclamatory words, 10) Words expressing feeling, 11) Word of calling,13) Words accepting calling

4.1. NOMINAL COMPLEXITY

Nominal forms show the following structure:

Noun (+Number) (+Case)

marng-kaL-ai ‘trees_PL_ACC’

Though at the underlying structure there are only two grammatical morphemes, on the phonological level, however, four types of morphs (or suffixes) can occur with the noun stem (Lehmann, 1989:12):

Plural suffix

Oblique suffix (increment)

Euphonic suffix (increment)

Case suffix

So we have the following structure (Lehmann, 1989:13)

Noun stem/Oblique stem (+euphonic increment) + case suffix)

Ex.

caavi.y-(in)-aal ‘key_euph_inst’

mara-tt(-in)-aal ‘tree_obl_euph_inst’

Noun stem + plural suffix (+euphonic increment) + case suffix)

Ex.

viiTu-kaL(-in)-ai ‘house_pl_euph_acc

Nouns need to be annotated into pronoun, proper noun and common noun. Pronouns need to be further annotated for person (1st, 2nd and 3rd), number (singular and plural), gender (masculine, feminine, neuter), status (honorific and non-honorific). Nouns need to be further distinguished into rational and irrational. Again nouns need to be annotated for nominative, accusative, dative, instrumental, sociative, locative, ablative, genitive, vocative cases.

Nouns and Pronouns need to be annotated as oblique or non-oblique form. In the following examples, *aaTu* is in nominative (i.e. non-oblique) form, where as *aaTTu* is oblique form; the formal difference makes difference in sense.

aaTu vs aaTTu ‘goat’

aaTu mandaiyil irukkiRatu ‘Goat is in herd’

aaTTu mandtaiyil avan oLindtaan

‘He hid himself in goat herd’

ndaan vs en ‘I’

ndaan en viiTTukkup pooneen

‘I my house_DAT go_PAST_FP

Further nouns need to be annotated for number and gender (masculine, feminine, and neuter) as the subject nouns show agreement with PNG marker at the finite verbal form. Nominaliztion makes the nominalized verbal form more complex. Nomininalized verbal forms need to be distinguished into two or three types. For example, Tamil requires the productive forms formed by the suffixation of tal/kai/aamai which are sentential in nature are to be differentiated from non-productive forms formed by the suffixation *ppu* etc. which are lexical in nature.

kaalaiyil ezhundu ndaTattal uTal ndalattiRku ndallatu

morning_IOC wake up_ADV walk_NOM health_DAT good

‘It is good to wake up and go for walking in the morning’

avan meel paTippu paTikka ameerikkaa cenRaan

he higher studies study_INF America go_PAS_he

‘He went to America for higher studies’

al-suffixed forms need to distinguished into two types as one type is lexical another is sentential.

(2) avaL aaTal avanaik kavrndtau

‘her dance he_ACC attract_PAS-he

‘He dance attracted him’

(4) avaL aaTal-aam ‘She may dance’

kai suffixed lexical form need be distinguished form its counterpart in sentential form.

vaazhkkai(il) ‘in life’ vs vaazhkai(il) ‘while living’

paTukkai is ambiguous

paTukkai(il) ‘in the bed’ / ‘while lying down’

atu suffixed ambiguous forms in Tamil need to be distinguished into three types.

vandtatu ‘that which came’/ ‘that somebody came’ /‘it came’

vandtat(u)-illai ‘did not come’

Many pronominalized forms are also ambiguous in Tamil and need to be distinguished into two types: lexical and sentential (productive).

paTittavan ‘educated male person’ / ‘one who read x’

4.3. VERBAL COMPLEXITY

The verbal forms are complex in Tamil. A finite verb shows the following morphological structure:

V+Tense+PNG

A number of non-finite forms are possible: adverbial forms, adjectival forms, infinite forms and nominalized forms

vandtu ‘having come’

varaamal/varaatu ‘without coming’

vara ‘to come’

vandtatu (illai)

Distinction need to be made between main verb followed by main verb and main verb followed by auxiliary verb.

Functional ambiguity in adverbial form

vandtu caappiTTu viTTu poo ‘having come and having eaten went’

ndondtu poo ‘become vexed’

Functional ambiguity in infinitival form

vara.v-iru ‘going to come’

vara-k.kuuTaatu ‘should not come’

vara-c.col ‘ask x to stop’

The adjectival forms differ by tense markings: V+Tense+Adjectivalizer

vandta ‘x who came’

varukiRa ‘x who comes’

varum ‘x who come’

Adjectival form allows several interpretations.

cappiTTa ilai ‘the leaf which is eaten by x’

‘the leaf on which x had his food and ate’

vaangkiya x ‘x which is bought’

‘x who bought’

‘x (price) by which something is bought‘

‘x (money) received’

‘x (container) in which something is received’

um-suffixed adjectival form clashes with other homophonous forms which leads ambiguity.

varum paiyan ‘the boy who will come’

varum ‘it will come’

varum pootu ‘while coming’

Some adjectivialized verbal forms are lexicalized adjectives (as against sentential ones)

iruNTa ‘dark’ (lexicalized) as in iruNTa kaalam ‘dark period’

iruNTa ‘dim’ as in (kaNNkaL iruNTana ‘eyes became dim’ >) iruNTa kaNkaL ‘the eyes which became dim’

ndiiNTa kai ‘long’ vs. ndiiNTa kai ‘extended hand’

Nominals which modify a noun function as adjectives.

mara-ppeTTi (T) ‘wooden box’

akala paatai ‘wide path

Verbal roots too functions as adjectives.

cuTu cooRu (T) ‘hot rice’

aazh kiNaRu (T) ‘deep well’

A number of adverbial forms of verbs functions as postpositions. They are discussed under ‘complexity in postpositions’.

C0MPLEXITY IN ADJECTIVES

Adjective are defined as a lexical category in Modern Tamil on the basis of the following syntactic properties: Adjectives occur as noun modifiers in a pre-nominal position between a determiner (Det) and the head noun (N).

inta nallap paiyan

When adjectives occur in predicate position, they have to be pronominalized.

inta payina nallavan ‘This boy is good’

Adjective formations

The past adjectival participle from the verb *aaka* ‘become’ ,that is *aana* occurs as a bound form added to nouns.

Oru azakaana peN ‘a beautiful girl’

aana has attained a suffix status and now it has be reanalyzed as a derivative suffix of denominal adjectives. It is not clear whether all *aana* suffixed forms can be called adjectives. Rajendran (2004) has listed 17 types of adjectives based on the entries in *Kriyavin taRkaalat tamiz akaraati*. They are listed below with a bit of statistical information.

Sr no	Types of adjective	Example	No of them found in KTTA
1	*aana* suffixed adjectives	*azakaana* 'beautiful'	100+
2	*enRa* suffixed adjectives	*jillenRa* 'cold'	
3	*cey* type of adjectives (the verbal root which are lexicalized as adjectives)	*aTar* 'thick'	17 +
4	*ceyta* type of adjectives (the past tense suffixed adjectival participle forms lexicalized as adjectives)	*akanRa* 'wide' (adjectival participle form of the verb *akal* 'leave' Similarly *aTarnta*, *iruNTa*, *izinta*, *keTTa*, *uTan piRanta*, etc.	117+
5	*ceykiRa* type of adjectives (present adjectival participle forms lexicalized as adjectives)	*talai pookiRa* (*pookiRa* is the adjectival present participle of the verb poo' go')	1+
8	*ceyyaa* type adjectives (negative participle forms without the final *ta* lexicalized as adjectives)	*iNai piriyaa*, *kaNkoLLaa* (*piriyaa* is the negative participle form of the verb *piri* 'separate' and *koLLaa* is the negative participle form the verb KoL 'have'	4+
9	The nouns which are lexicalized as adjectives	talaimai ' tiruTT, tuNai	22+
10	The nouns with the final doubling of consonant lexicalized as adjective	*veeRRu* (from noun veeRu *naaTTu* (from noun naaTu)	5+
11	The nouns with the final m eluded lexicalized as adjective	*akila* (from noun *akilam*), *kaLLa* (from *kaLLam*), *cutta* (from *cuttam*), *avacara* (from avacaram '	42+

5	***ceykiRa* type of adjectives (present adjectival participle forms lexicalized as adjectives)**	***talai pookiRa* (*pookiRa* is the adjectival present participle of the verb poo' go')**	**1+**
6	***ceyyum* type of adjectives (future adjectival participle forms lexicalized as adjectives)**	***nikazum* 'pesent', *varum* 'future' (*nikazum* is the future adjectival participle form of the verb *nikaz* 'happen' and, *varum* 'future' (the future adjectival participle form of the verb *vaa* 'come') Similarly *naTamaaTum*, *pinRRaum,kiizkkaaNum***	**7+**
7	***ceyaata* type of adjectives (negative adjectival participle forms lexicalized as adjectives)**	***teriyaata*, *veNTaata* (*teriyaata* is the negative adjectival participle form of the verb teri 'know' and *veeTaata* is the negative adjectival participle form of the verb *veeNTu* 'request'. Similarly iTaiviTaata, iNai piriyaata, vaRalaaRu kaaNaata, etc.**	**22+**

No

Nouns used as adjectives

We have noticed that many nouns have been used as modifiers of another head noun. So nouns attributing another noun naturally becomes adjectives at least function wise.

marap peTTi 'wooden box'

irumpuk katavu 'metal door'

tangka naaNayam 'gold coin'

All the colour terms are functionally adjectives; their adjectival forms are bound forms.

paccai – pacum 'green'

civappu – cem

veLLai - veN

kaRupuu – karum

Nouns used or lexicalized as adjectives

As we have noticed many nouns ending in *am* (e.g. *maram* ‘tree/wood’, *tangkam* ‘gold’) attributing a head noun functions as adjectival modifiers. Some of them are lexicalized as adjectives. Sometimes they are falsely interpreted as adjectives as they have a-ending form. Such N+N combination is productive and predictable. So one may be tempted to put such attributive nouns as adjectives. Sometimes we are forced to reinterpret them as adjectives when we work across languages (for example for bilingual dictionaries, translational equivalents. Many are tempted to annotate *peN* of *peN cingkam* as adjective.

cey type of nouns used as adjectives

The presence of cey type of adjectives may lead to misinterpretation. cuTu of cuTu cooRu ‘hot rice’ and tuungku of tuungku muunjci ‘dull person’ can cause misinterpretation.

The lexicalization of all types of adjectival participle forms (ceyta, ceykiRa, ceyyum, ceyyaata, ceyyaa types of adjectival participle forms) may cause ambiguity in the annotation of adjectives. For example the adjectival participle form teernta is lexicalized as an adjective.

Problem of bound adjectives

Another question often raised to me is whether to split the bound forms of adjectives from the head noun or not. The situation is similar to black bird; where we often wonder whether *black bird* is a type of bird of a bird which is black. Naturally in Tamil for example we wonder whether to split *centaamaramai and veNtaamarai* into *cem* and *taamarai* and *veN + taamarai* and annotate them accordingly. As they are mostly interpreted as compounds they are left as full nominal forms.

Problem with _aana_ suffixed forms

There are *aana* suffixed forms which we find difficult to accept at adjectives. For example in the phrases *aaciriyaraana raaman* ‘teacher Raman’, *maaNaviyaana celvi* ‘student Selvi’, are we going to annotate *aaciriyaraana* as adjective or not?

Demonstrative

There are two types of determiners. Both are demonstratives, which occur in pre-nominal position. They specify or identify the referent’ proximity to the speaker. The proximate demonstrative determiner is *inta* ‘this, these’ and the remote demonstrative is *anta* ‘that, those.

Proximate and remote demonstrative

Sr. no.	Demonstrative	Words	Examples
1	Proximate Demonstrative	Inta 'this, these'	Intap paiyan 'this boy Inta paiyankaL 'these boys'
2	Remote Demonstrative	Anta 'that, those'	anta viiTu 'that house' anta viiTukaL 'those houses'

Problem of bound demonstrative

So far it is fine. But we have bound demonstrative forms which are prefixed to the nouns. They create problem in annotation. i is the bound form of inta and a is the bound form of anta. They need to be split from their respective nouns.

i-p-paiyan 'this boy

a-p-paiyan 'that boy'

But a- is homophonous with the negative prefix borrowed from Sanskrit.

a-niyaayam 'injustice'

a-naakariikam '

4.4. COMPLEXITY IN ADVERBS

We have seen that a number of adjectival and adverbial forms of verbs are lexicalized as adjectives and adverbs respectively and clash with the respective sentential adjectival and adverbial forms semantically creating ambiguity in POS tagging. Adverbs too need to be distinguished based on their source category. Many adverbs are derived by suffixing *aaka* with nouns in Tamil. But not all *aaka* suffixed forms are adverbial.

veekam-aaka (T) 'fast' vs. TaakTar-aaka 'as doctor'

Functional clash can be seen between adjective and adverb in *aaka* suffixed forms. This type of clash is seen among other Dravidian languages too.

avaL azhkaaka irukkiRaaL

‘she beauty_adv be_pre_she

‘she is beautiful’

The homophonous lexicalized and non lexicalized adverbs created ambiguity.

avan paarttu vandtaan

he see_ADV come_PAST_he

‘He came having seen’

‘He came carefully’

avan viraindtu vandtaan

‘he came fast’

4.6. COMPLEXITY IN POSTPOSITIONS

Postpositions are from various categories such as verbal, nominal and adverbial in Tamil. Many times the demarking line between verb/noun/adverb and postposition is slim leading to ambiguity. Some post positions are simple and some are compound. Postpositions are conditioned by the nouns inflected for case they follow. Simply tagging one form as postposition will be misleading.

There are postpositions which come after noun and also after verbs which makes the postposition ambiguous (spatial vs. temporal).

pinnaal ‘behind’ as in viiTTukkkup pinnaal ‘behind the house’

pinnaal ‘after’ avanukkup pinnaal vandtaan ‘he came after him’

Use of adverbial forms of verbs leads to ambiguity in the annotation of postpositions,

katti koNTu

knife have_ADV

‘by means of knife/having the knife’

avaLaik koNTu

she_ACC have_ADV

‘by means of her/having her’

Similarly the following adverbial forms leads to problems in POS tagging.

viiTTil irundtu

house_LOC be_ADV

‘from the house/being in the house

The complex postpositions still makes things more complex.

viiTT-il-irundtu-koNTu

house_LOC be_ADV_have_ADV

‘being at home’

Similarly the following postpositions from verbal source (adverbial forms of certain verbs may lead to ambiguous annotation.

oTTi ‘regarding’ the adverbial from of the verb oTTu ‘stick

kuRittu ‘about’ the adverbial form of the verb kuRi ‘aim, mark’

cuRRi ‘around’ the adverbial form of the verb cuRRu ‘circulate’

tavirttu ‘except’ the adverbial form of the verb tavir ‘avoid’

paRRi ‘about’ the adverbial form of the verb paRRu ‘seize’

viTTu ‘from’ the adverbial form of the verb viTu ‘leave’

vaittu ‘with’ the adverbial form of the verb vai ‘put’

ndokki ‘towards’ the adverbial form of the verb ndookku ‘see’

pindti ‘after’ the adverbial form of the verb pindtu ‘be behind’

mundti ‘before’ the adverbial form of the verb mundtu ‘precede’

tavira ‘except’ the infinitive form of the verb tavir ‘avoid’

viTa ‘than’ the infinitive form of the verb viTu ‘leave’

Tamil is no doubt a morphologically rich language. The relation between verb and its nominal arguments are decided by case suffixes rather than position. It is possible to have a few number of tagset at shallow level. But one need to address other unique features at the deep level. Hierarchical tagset is a welcome thing.

3. CHUNKING

Various Natural Language Processing (NLP) tasks can be seen as a chunking task. Machine learning methods are often applied to chunking, since the task is viewed as approximating and identifying function from the information available in the adjacent context. Various machine learning approaches have been proposed for chunking. The chunking cover up many different areas of research and the concept of a chunk has accordingly diversified in its meaning. Chunk parsing, also called shallow parsing or partial parsing, is a technique described by Steven Abney (1991). Chunking can be seen as the basic task in fractional Parsing. Fractional Parsing was introduced as a response to the difficulties of the full traditional parsing and it is described as a technique to recover syntactic information efficiently and reliably from unrestricted text, by sacrificing completeness and depth of analysis. Chunking or shallow parsing segments a sentence into a sequence of syntactic constituents or chunks, i.e. sequences of adjacent words

grouped on the basis of linguistic properties. Chunking or local word grouping refers to the identification of syntactically correlated parts of words in a sentence and is usually the first step towards parsing of a natural language sentence. This process can be carried out efficiently and thus chunking can be useful in several tasks, for instance Terminology Discovery, Named Entity Recognition and Text Mining, and also as an intermediate step providing input to further full parsing stages.

Chunking plays an important role in parsing, information extraction and information repossession. Chunking is often a useful preprocessing step for many bilingual tasks, such as machine translation, cross language information retrieval, etc. A chunk is the largest sub graph of a major head "that does not contain the root of another chunk". A major head is a semantic concept. Chunking aims at labeling segments of a sentence with syntactic constituents such as noun or verb phrases (NP or VP). Each word is assigned only one unique tag. Every chunk can be retrieved by a single act of identification.

Abney introduces the concept of chunks in the field of NLP and parsing, motivated from psychological as well as phonological evidence presented in his paper. He defines chunks as non-recursive, non-overlapping constituents with exactly one main head, where a main head is a content word that is not between a function word and the content word it selects. Education and cognition has uncovered substantial evidence for a unifying information-processing mechanism, known as 'chunking'. Miller's key contribution was to propose an information gauge for cognitive systems based upon the concept of a 'chunk', where each chunk collects a number of quanties of information from the environment into a sole unit.

The chunker is divided into three levels. In the first level most of the lexical information is established, such as e.g. lexical or semantic features. The second level is the main chunking level. In this level, the word forms and pos tags (Part of Speech Tagging) can be almost fully ignored since the rules now refer to structures built in the first level or in previous iterations of the second level. The third level has mainly two tasks. First it checks for correct structures and rejects, if necessary, some of the imaginary structures built in the second level. Secondly it builds the hierarchical structure of the chunks.

Chunking is a kind of human language processing. There are psycholinguistic evidences or arguments for chunking process. For example, when a person reads through a text, he or she may not fully parse the sentence but instead look for "key phrases". These key phrases are actually chunks, which means that chunking is the machine analogy for speed reading .When a speaker of a language reads a sentence, the total sentence's prosody features, intonation changes and pauses, often match the edges of chunks as they would have been output by a shallow parser.

The chunk types are based on the syntactic category part of the bracket label in the Treebank. Roughly, a chunk contains everything to the left of and including the syntactic head of the constituent of the same name. Some Treebank constituents do not have related chunks. The head of S (simple declarative clause) for example is normally thought to be the verb, but as the verb is already part of the VP chunk, no S chunk exists in our example sentence. Besides the head, a chunk also contains pre-modifiers (like determiners and adjectives in NPs), but no post-modifiers or arguments. This is why the PP chunk

only contains the preposition, and not the argument NP, and the SBAR chunk consists of only the complementizer.

CHUNK PARSING

A chunk is a textual unit of adjacent word tokens which can be mutually linked through unambiguously identified dependency chains with no recourse to idiosyncratic lexical information. Pisa and Stuttgart have developed somewhat varying notions of what a chunk is, aimed at somewhat different goals. The standard notation that we present here is the result of a joint effort aimed at evaluation and comparison of the two proposed annotation schemata, and covers their proper intersection. Aspects specific to either representation schema are dealt with here only insofar as they are considered relevant to the evaluation phase.

Given the sentence *the interested watcher could always observe the stars visible to the naked eye*, the chunking will be as follows:

[*the interested watcher*] [*could always observe*] [*the stars*] [*visible*] [*to the naked eye*] The sentence is segmented into five chunks. Each chunk includes a sequence of adjacent word tokens (a text substring) which are mutually related through dependency links. For example, figure 1 shows the dependency chains within chunks -4 and -3.

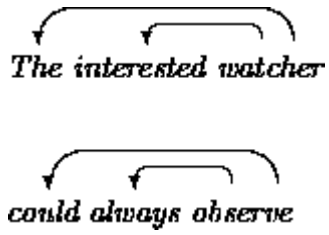

Abbildung 1: Dependency links within chunks

The fact that two text substrings are assigned to different chunks does not necessarily entail that there is no dependency relationship linking the two. Simply, on the basis of the available

knowledge it is impossible to state unambiguously what chunk relates to its neighbouring chunks and what the nature of this relationship is. For example, if we abstract away from the lexical content of chunks, the final chunk in the above sentence is potentially dependent on chunks -1, -2 or -3 and this cannot be decided unless lexical knowledge of some kind (e.g. that *visible* subcategorizes for a prepositional phrase headed by *to*) is resorted to.

All this amounts to saying that chunks are defined strictly syntactically: following the definition of Abney (in press), a chunk can be conceived of as ``the non-recursive core of an intra-clausal constituent, extending from the beginning of the constituent to its head, but not including post-head dependents". Examples and criteria for chunking are given in the following sections.

Steve Abney pioneered the notion of chunk parsing.

[I begin] [with an intuition]: [when I read] [a sentence], [I read it][a chunk] at a time.

Chunks appear to correspond roughly to prosodic phrases (though nobody has really researched this form the phonology-syntax interface perspective). Strong stress will be once a chunk. Pauses are most likely to fall between chunks. Chunks are difficult to define precisely. “A typical chunk consists of a content word surrounded by a constellation of function words, matching a fixed template.” (Abney 1991).

[I begin] [with an intuition]: [when I read] [a sentence], [I read it] [a chunk] [at a time]

[The girl] saw [the monkey] [with the telescope]

By contrast, the relationship between chunks is mediated more by lexical selection than by rigid templates.”(Abney 1991).

Abney envisioned chunk parsing as cascade of (finite-state) processes:

Word Identification (in practice mostly done via POS tagging)

Chunk Identification (via rules)

Attachment of Chunks (Parsing, also via rules)

Basically, a chunker divides a sentence into its major non-overlapping phrases and attaches a label to each.

> [NP The student] from [NP Germany] [VP was trying to phone] [NP his parents] [NP last night]

Chunkers differ in terms of their precise output and they way in which a chunk is defined. Many do more than just simple chunking. Others just find NPs. Chunking falls between tagging (which is feasible but sometimes of limited use) and full parsing (which is more useful but is difficult on unrestricted text and may result in massive ambiguity). The structure of individual chunk is fairly easy to describe, while relationship between chunks are harder and more dependent on individual lexical properties. So chunking is a compromise between the currently available and the ideal processing output. Chunkers tokenise and tag the sentence. Mostly chunkers simply use the information in tags, but other look at actual words.

3.1. Discussion on chunking and POS tagging stardards

a) Chunking Standards

b) POS Tagging Standards (a revisit)

c) Dictionary Standards

Various issues related to the above three points were discussed extensively during the two day meeting. Sections 2, 3 and 4 of this document contain a detailed account of the deliberations, decisions, conclusions on chunking standards, POS Tagging standards and dictionary standards respectively.

3.1.1. Chunking Standards

Various points which were deliberated upon in relation to chunking scheme were :

1) Definition of a chunk
2) Chunk Types
3) Annotation method/procedure

3.1.1.1. Definition of a chunk

Following issues related to the definition of a chunk were discussed :

What constitutes a ‘chunk’?

Chunks are normally defined as 'correlated chunks of words' (Steven Abney,1991). The question then is,

How to define the boundaries of these 'correlated word groups’?

((xillI meM)) or ((xillI)) meM
((rAjA kA betA)) or ((rAjA kA)) ((betA))
((rAjA ke bete kI paxnI)) or ((rAjA ke)) ((bete kI)) ((paxnI))

Following definition of a 'chunk' was evolved through discussion:

“A minimal (non recursive) phrase(partial structure) consisting of correlated, inseparable words/entities, such that the intra-chunk dependencies are not distorted”.

3.1.1.2. Chunk Types

Based on the above definition of chunk, issues related to various chunk types were discussed. A chunk would contain a 'head' and its modifiers.

Chunk Tagset

The following could be the list of possible chunks.

Main Tagset		Representation
1	Noun Phrase Chunk	NP
2	Verb Finite Chunk	VGF
3	Non-Finite Verb Chunk	VGNF
4	Conjunction Chunk	CCP

5	Verb Chunk Gerund	VGNN
6	Negation Chunk	NEGP

1. **NP** - Noun Chunks will include minimal noun phrases and prepositional phrases. The head of a noun chunk would be a noun. Elements other than noun which can be part of a noun chunk are determiners, adjectives, quantifiers etc before the head noun and post positions (PREP) after the noun.
2. **VG** - A verb group will include the main verb and its auxiliaries, if any.
3. **JJP** - An adjectival chunk will be tagged as JJP. This phrase will consist of all adjectival chunks including the predicative adjectives. However, adjectives appearing before the noun will be grouped together with a noun chunk.
4. **RBP** - This chunk name is again in accordance with the tags used for POS tagging. This chunk will include all pure adverbial phrases.
5. **CCP** -It is important to annotate conjunct info separately at various levels. Conjuncts are functional units which information about which is required to build the larger sentence structures. Conjunctions can consist of more than one word as well: *as well as, instead of, rather than, not only, but also.*
6. **INTJ -** It is an interjection phrase/chunk like *no, oh, hello, alas, good grief!*. It is quite rare.
7. **UCP -** This chunk is reminiscent of the UCP (unlike coordinated phrase) constituent in the Treebank. Arguably, the conjunction is the head of the UCP, so most UCP chunks consist of conjunctions like *and* and or. UCPs are the rarest chunks and are probably not very useful for other NLP tasks.

3.1.1.2.1 Noun Chunks (NP)

Noun Chunks will be given the tag NP and include non-recursive noun phrases and postpositional phrases. The head of a noun chunk would be a noun. Specifiers will form the left side boundary for a noun chunk and the vibhakti or head noun will mark the right hand boundary for it. Descriptive adjective/s modifying the noun will be part of the noun chunk. The particle which anchors to the head noun in a noun chunk will also be grouped within the chunk. If it occurs after the noun or vibhakti, it will make the right boundary of the chunk.

Some example noun chunks are :

((bacce_NN))_NP, ((kucha_QF bacce_NN))_NP,

'children' 'some' 'children'

((kucha_QF acche_JJ bacce_NN))_NP, ((Dibbe_NN meM_PSP))_NP,

'some' 'good' 'children' 'box' 'in'

((eka_QF kAlA__JJ ghoDZA_NN))_NP ,

'one' 'black' 'horse'

((yaha_PRP nayI_JJ kitAba_NN))_NP,

'this' 'new' 'book'

((isa_PRP nayI_JJ kitAba_NN meM_PREP))_NP,

'this' 'new' 'book' 'in'

((isa_PRP nayI_JJ kitAba_NN meM_PREP bhI_RP))_NP

'this' 'new' 'book' 'in' 'also'

The issue of genitive marker and its grouping with the nouns that it relates was discussed in detail. For example, the noun phrase 'rAma kA beTA' contains two nouns 'rAma' and 'beTA'.

The two nouns are related to each other by the vibhakti 'kA'. The issue is whether to chunk the two nouns separately or together. Linguistically, 'beTA' is the head of the “rAma kA beTA”.

'rAma' is related to 'beTA' by a genitive relation which is expressed through the vibhakti 'kA'. Therefore, going by our definition of a 'chunk' we should break 'rAma kA beTA' into two chunks by breaking 'rAma kA' at 'kA' vibhakti. Moreover, if we chunk 'rAma kA beTA' as one chunk, linguistically, we will end up with a recursive noun phrase as a single chunk ((((rAma kA)) beTA)) which also is against our definition of a chunk.

Decision taken: Genetive markers will be chunked along with the preceeding noun. Therefore, it was decided to chunk the noun group 'rAma kA beTA' into two chunks

eg. **((rAma kA))NP** ((beTA))NP

((kitAba))NP **((rAma kI))NP** hE

However, noun group such as “usakA beTA” will be chunked together for Hindi.

3.1.1.2.2 Verb Chunks (VG)

The tag for verb chunk would be VG. A verb group will include the main verb and its auxiliaries, if any.

eg. ((khAyA))_VG, ((khA rahA hE))_VG, ((khA sakawe hEM))_VG etc.

3.1.1.2,3 Conjuncts

The special handling of conjunct verb constructions in ILs was discussed.

The issue whether to treat the noun/adjective which is part of a conjunct verb differently by marking it with a special tag (NVB/JVB) or to treat it as a noun like any other noun at the POS level was discussed during the earlier workshop on POS tag standards. The issue was discussed further here for the chunk level.

The question was based on the following observations _

a) NVB/JVB , as part of conjunct verbs, are most often not recognized by the learning

algorithms.

b) Having NVB at the POS level is based on syntactic considerations. Therefore, do we really need to go for it ? Instead, at the POS level we mark the noun as a noun and leave the decision of marking a conjunct verb as single unit for a later level.

c) Also, since the noun which is part of a conjunct verb (Kriyamula) can also occur away from its 'verbaliser', it becomes difficult to differentiate it from a 'noun' participant. This also creates problem for chunking of the VG. The two components of the chunk have to be separately marked and have to be joined at the syntactic level.

d) If NVB is marked at the POS level, a natural consequence would be to group it with its verbaliser as a VG chunk. In fact, that is the purpose of identifying it as different from a noun. However, sometimes one comes across expressions such as 'mEMne unase eka prashna pUCA'. In this sentence, 'eka' is a modifier of 'prashna'. 'prashna karana' is recognized as a conjunct verb in Hindi by most Hindi speakers. Following example shows the problem of grouping 'praSna karanA' as a single VG:

POS : mEMne_PRP unase_PRP eka_QC prashna_NVB kiyA_VF
Chunk: ((meMne_PRP))_NP ((unase_PRP))_NP ((eka_QC))_JJP ((prashna_NVB kiyA_VF))_VG

If we group “praSna karanA” as a chunk, it is difficult to show the relation between 'eka' and 'prashna' .

Thus, an alternative was proposed wherein, the noun of the conjunct verb is tagged as NN at the POS level. Thereafter, the noun is grouped with its preceding adjectival modifiers as an NP chunk. The only problem in this approach is that the information of a noun verb sequence being a 'kriyamula' is not captured till the chunk level and the noun of the 'kriyamula' is separated from its verbaliser. However, the approach has following advantages:

1) At the POS level, the word is tagged for its grammatical category and not for its syntactic function. This eases the decision making at the POS level. And marking the information, that the conjunct verbs which are composed of two words form one lexeme semantically, is postponed to a later level.

2) It allows us to show the modifier-modified relation between an adjective such as 'eka' in the above example with its modified noun 'praSna'.
3) Since the information of a noun verb sequence being a 'kriyamula' is crucial at the syntactic level, it will be captured at thatlevel by marking the relation between the 'noun' and its verbaliser by an appropriate tag. Therefore, the decision is:
The noun/adjective and verb (internal components of a conjunct verb) will be chunked separately.

eg. **((prashna))NP ((karaA))VG**

((ucita))JJP ((kiyA))VG

3.1.1.2.4. Some special cases

Notice the following examples -

a) dillI meM ((rahanevAlA))_VG merA bhAI.

b) usane Tebala para ((rakhA huA))_VG seba khAyA.

In (a) above 'rahanevAlA' is an ajectival participial. But we do NOT mark it as JJP. Instead, it will be marked as a VG. The decision to tag it as a VG is based on the fact that such participials are derived from a verb and the verb retains some of its verbal properties. Therefore, it is important to preserve the information that the word is a form of verb and is capable of taking its own arguments in a sentence. This information is useful for processing at the syntactic level. Similarly, in (b) above the chunk 'rakhA huA' is an adjective but will be marked as a VG as it is a form of verb and chunks like 'Tebala para' etc are its arguments. So, it will be treated as a verb chunk and marked as VG.

Similarly, in the examples below, 'dagamagAte hue' and 'khAkara' are non finite forms of verbs used as adverbs. They will also be chunked as VG and not as RBP.

c) vaha ((dagamagAte_VNF hue_VAUX))_VG cala rahA thA.

d) vaha khAnA ((khAkara_VNF))_VG ghara gayA.

3.1.1.2.3 Adjectival chunk (JJP)

An adjectival chunk will be tagged as JJP. This phrase will consist of all adjectival chunks including the predicative adjectives. However, adjectives appearing before the noun will be grouped together with a noun chunk. A JJP will consist of phrases like

1. vaha laDZakI hai ((suMdara sI))_JJP

2. hAthI AyA ((moTA tagadA))_JJP

However, cases such as 3 below will not have a JJP. In these cases the adjectives will be grouped together with the noun they modify, thus forming a NP chunk.

3. ((kAle ghane laMbe bAla))_NP

3.1.1.2.4 Adverb chunk (RBP)

This chunk name is again in accordance with the tags used for POS tagging. This chunk will include all pure adverbial phrases.

a) vaha **((dhIre-dhIre))_RBP** cala rahA thA.

3.1.1.2.5 Conjuncts (CCP)

Conjuncts are functional units information about which is required to build the larger structures. Take the following examples of cunjunct usages:

1. rAma kitAba paDha rahA thA Ora mohana tennisa khela rahA thA.

2. rAma ne batAyA ki usakI kitAba acchI hE.

3. rAma Ora mohana tennisa khela rahe the.

4. merA bhAI rAma Ora usakA dosta mohana tennisa khela rahe the.

5. rAma saphZeda kapade Ora nIle jute pahane thA.

6. rAma sAphZa jute Ora safeda kapade pahane thA.

7. rAma eka halkI Ora nIlI bOla lAyA.

The sentences above have various types of conjoined structures. To represent these conjoined structures, it is decided to form separate chunks for conjuncts and the elements a conjunct conjoins. Thus 1 and 2 above will be chunked as 1(a) and 2(a) given below,

1a. ((rAma))_NP ((kitAba))_NP ((paDha rahA thA))_VG **((Ora))CCP** ((mohana))_NP ((tennis))_NP ((khela rahA thA))_VG.

2a. ((rAma ne))_NP ((batAyA))_NP **((ki))_CCP** ((usakI))_NP ((kitAba))_NP ((acchI))_JJP ((hE))_VG.

Expression 'rAma Ora mohana' in example 3 is a complex NP. Though complex, the expression can be annotated as a single NP chunk as fucntionally it is the subject of the verb 'play'. However, example 4 presents a case where it would be better to form three independent chunks for the complex subject NP. Though the conjunct 'Ora' is conjoining 'rAma' and 'mohana', both 'rAma' and 'mohana' have their respective modifiers. To make it explicit, it is better to treat them as two independent NP chunks conjoined by a CCP.

4a. ((merA bhAI rAma)) **((aura))_CCP** ((usakA dosta mohana))_NP ((tennis))_NP ((khela rahe the))_VG.

Following this, the subject NP of 3 would also be annotated similarly. Therefore,

3a. ((rAma))_NP **((aura))_CCP** ((mohana))_NP ((tennis))_NP ((khela rahe the))_VG.

The annotation for cases such as 5, 6 and 7 would be as follows :

5a. ((rAma))_NP ((safeda kapade))_NP ((aura))_CCP ((nIle jute))_NP ((pahane thA))VG.

6a. ((rAma)) ((sAfa))_JJP ((aura))_CCP ((safeda kapade))_JJP ((pahane thA))_VG.

((rAma))_NP ((eka))_JJP ((halkI))_JJP ((aura))_CCP ((nIlI))_JJP ((bOla))_NP ((lAyA))_VG

Thus the decision for conjuncts is :

The conjoined entities will be broken into separate chunks.

eg. ((rAma))NP **((Ora))CCP** ((SyAma))NP

3.1.1.2.6 Negatives (NEGP)

(i) In case a negative particle occurs around a verb, it is to be grouped within verb group. For example,

mEM kala dillI **nahIM jA rahI**

"I" "yesterday" "Delhi" "not" "go" "Cont"

((meM))_NP ((kala))_NP **((nahIM jA rahI))_VG**

Also,

binA bole kAma **nahIM calatA**

"without" "saying" "work" "not" "happen"

((binA bole))_VG ((kAma))_NP **((nahIm calatA))_VG**

However ,

binA kucha **bole** kAma **nahIM calatA**

"without" "something" "saying" "work" "not" "happen"

In the above sentence, the noun "kucha" is coming between the negative "binA' and verb "bole". Here, it is not possible to group then negative and verb as one chunk. At the same time, "binA" cannot be grouped within an NP chunk as functionally, it is negating the verb and not the noun. To handle such cases an additional NEGP chunk is introduced which will be used only if the NEG occurs away from the verb. Therefore,

(ii) If it is away from the verb/noun it is marked as a separate chunk.

((binA))_NEGP ((kucha))_NP **((bole))_VG ((**kAma))_NP ((**nahIM calatA))_VG**

3.1.1.2.7. Chunk Fragments (FRAGP)

Some fragments of chunks are separated from the chunk to which they belong. For example :

1) **rAma** (jo merA baDZA bhAI hE) **ne** kahA

In the above example vibhakti 'ne' which is a marker of the noun 'rAma' is separated from it by an intervening clause. Syntactically, 'ne' is a part of the noun chunk 'rAma ne', however, it can be written separately. The following was decided for such fragments :

(i) There will be a separate chunk for the vibhakti in constructions where it gets separated from the noun it would normally be grouped with. This chunk can have more than one entity within it.

eg. ((**rAma))** (merA dillI vAlA bhAI) **((ne)) FRAGP** kahA

((**rAma))** (merA dillI vAlA bhAI) **((ne)) FRAGP** kahA

(ii) If the entities embedded between the noun and it's vibhakti is a series of nouns the entire group will be chunked as one noun chunk.

eg. ((isa 'upanyAsa samrATa' Sabda kA)) NP

3.1.1.2.8 Miscllaneous entities (BLK)

Entities such as interjections and discourse markers that cannot fall into any of the above mentioned chunks will be kept within a separate chunk.

eg. ((oh)) BLK

((arre))BLK

3.1.1.2.3 Some more issues

Apart from the above, some more issues related to certain lexical types were also discussed.

3.1.1.2.3.1 Particles

Regarding the particles it was decided that the particles will be chunked with the same chunk as the anchor word it occurs with.

eg. ((rAma ne **bhI**))**NP**

3.1.1.23..2 Quantifiers

The issue of chunking quantifiers was discussed in great details. Numbers can occur (a) as noun modifiers before a noun (haZaroM ladakoM ne) or (b)can occur without a noun (hazAroM ne) with a nominal inflection. The issue of whether to treat the quantifiers of the type (b) as nouns was discussed whether (b) is a case of an ellipsis of the noun after a number or is the number itself which is the noun. If the latter has to be followed then the POS tag for quantifiers in such cases should be NN. Following decisions were taken :

(i) A 'QC' or 'QO' occuring with a noun will be part of the noun chunk.

eg. ((hazAroM/QC logoM/NN ne)) NP

((dUsare ladake ne)) NP

(ii) All categories occuring without a noun, with nominal inflections (overt or otherwise) will be tagged as noun.

eg. ((**hazAroM/NN** ne)) NP

((mote/NN ne))

3.1.1.2.3.3. Punctuation

Decision taken:

All punctuations, with an exception of sentence boundary markers and clausal conjuncts will be included in the preceding chunk.

eg. ((usane)) NP ((**kahA –)**)VG (("**yaha**)) NP ((Thika))JJP ((**hE** "))VG

eg. rAma AyA **((,))CCP** mohana gayA

Punctuations such as (a) hyphens and (b) quote marks will be taken care of by the tokenizer.

(a) Hyphens: Identified to be of two types:-

(i) without space on either sides, as in the case of compound nouns

eg. mAtA-pitA

(ii) with spaces, as in the case of rAma ne kahA – yaha thIk hE

(b) Quote marks (single and double both) : Identified to be of two types:-

(i) opening

(ii) closing

3.1.1.2Annotation Procedure

To maintain consistency in data format and the annotation, it was decided to use 'Sanchay', a facility developed at IIIT, Hyderabad for the annotation task.

4. Conclusion

The discussion on annotation standards for Indian languages concluded with a decision on the following chunk types for the POS/chunk annotation task

NP	noun chunk	((merA nayA ghara))
VG	verb chunk	((nahIM jA rahA hE))
JJP	adjective chunk	((bahuta acchA))
RBP	adverb chunk	((bahuta jaldI))
CCP	conjunct chunk	((Ora)), ((yA)), ((ki))
NEGP	negation chunk	((binA)), ((nahIM))
FRAGP	fragment chunk	((**rAma))** (merA dillI vAlA bhAI) **((ne))** FRAGP
BLK	miscellaneous	((,_SYM)) BLK

Amrita Chunk set

This POS tagged corpus is used for chunking corpus development. Our customized chunk Tagset was used to tag the chunking corpus of size two lakh and twenty five thousand words. See fig.2.

S.NO	ChunkTag	Tag name	Possible POS Tags
1	NP	Noun Phrase	NN,NNP,NNC,NNPC,NNQ,PRP, QTF,DET,CRD,ORD,ADJ,INT
2	AJP	Adjectival Phrase	NN,ADJ
3	AVP	Adverbial Phrase	ADV,INT,CRD
4	VFP	Verb Finite Phrase	VF,VAX
5	VNP	Verb Nonfinite Phrase	VNAJ, VNAV,VINT,CVB
6	VGP	Verbal Gerund Phrase	VBG
7	CJP	Conjunctional Phrase	CNJ
8	COMP	Complimentizer	COM
9	O	Symbols	. ?

REFERENCE

Abney. S .1991. “Parsing by Chunks”. In Principle-Based Parsing. Kluwer Academic Publishers, Dordrecht.

Akshar Bharati, Rajeev Sangal, Dipti M Sharma 2007. SSF: Shakti Standard Format Guide Language Technologies Research Centre International Institute of Information Technology Hyderabad, India {sangal,dipti}@iiit.ac.in, 30 September 2007

AU-KBC POS TAGSET for TAMIL. AU-KBC Research Centre,MIT Campus of Anna University, Chromepet, Chennai-44.

EAGELES Recommendations for the Morphostyntactic Annotation of Corpora. EAGLES Document EAG-TCWG-MAC/R Version of Mar, 1996.

Unified Parts of Speech (POS) Standard in Indian Languages - Draft Standard –Version 1.0. Department of Information Technology Ministry of Communications & Information Technology Govt. of India.

AnnCorra : Annotating Corpora Guidelines For POS And Chunk Annotation For Indian Languages Akshar Bharati, Dipti Misra Sharma, Lakshmi Bai, Rajeev Sangal Language Technologies Research Centre IIIT, Hyderabad, Date : 30-11-2006

Chunking Annotation Standards, LTRC, IIIT Hyderabad , Oct 6-7, 2006

ILMT System Software Requirement Specifications (SRS) by ILMT Consortium Draft January-2007 IIIT Hyderabad Gachibowli, Hyderabad 500032.

Abney, S.1991. R. Berwick, S. Abney and C. Tenny (eds), "Parsing by chunks in Principle-Based Parsing", Kluwer Academic Publishers, Dordrecht.